THE UNIVERSITY
of EDINBURGH

A Novel Role of an ER-Resident Chaperone Pathway in Cancer Signalling

Mohamad Aimanuddin Mohtar

THE UNIVERSITY *of* EDINBURGH

Doctor of Philosophy

The University of Edinburgh

2017

Declaration

I hereby declare that I am the author of this thesis. The work herein is entirely my own unless otherwise clearly indicated and acknowledged. I can confirm that this thesis has been submitted for the degree of Doctor of philosophy and no part of this work has been submitted for any other degree or professional qualification.

Mohamad Aimanuddin Mohtar

Acknowledgements

My most humble and genuine thanks to my supervisor Professor Ted Hupp; for giving me the opportunity to join your team; for your endless support and advice; for being the coolest and philosophical supervisor; and for always saying 'encouraging' results in otherwise disastrous experimental results.

Special thanks to my extraordinary wife, Zulaikha, for her love, patience, good food and cakes during my entire PhD journey and for listening to my bellyaches. To my beloved, most precious Scottish-born daughter, Zahra who came into the world at the beginning of my PhD, cheering me up every day and growing up stronger parallel to my PhD thesis.

Many thanks to the entire team of *Cell Signalling Unit* (Prof Ted Hupp and Prof Kathryn Ball) for being supportive, friendly, and helpful - Ali, Bhindu, Maria Gil, Jia, Jonas, Luke, Maria Gomez, Ashita, Fiona, Erisa, Lucy, Clare and Catarina. To the 'AGR2 team' – Stephanie, Puay and Srividya for making me feel that I'm not alone. Much great appreciation also to the past members of the lab especially to Dr Euan for lab tips, protocols and more importantly, your AGR2 stuff giveaways, to Dr Saurabh, Dr Jianguo, Dr Liz, Dr Kalai, Li R and Elisabeth.

My sincere gratitude to the collaborators in Brno, Czech Republic – Dr Lenka Hernychova and Dr Borek Vojtesek for mass spectrometry analysis. Also thanks to the imaging specialists Matt, Harris, and Dr Ann at IGMM imaging facilities for making me a good epifluorescent and confocal microscopist and to Dr Alison Dun and Prof. Rory Duncan at Heriot-Watt University for help with TIRF microscopy and live-cell imaging. To the procurement and support team Jim, Gary, Derek, Jade and Shirley for making sure we have everything we needed in the lab to do amazing *Science*.

Many thanks also to my fellow homies - Emran, Aizi, Anas and everyone in the *EMSAF* for good cheer and making me feels like I'm closer to home, to Mushadad for your warm welcome when we first arrived in Edinburgh.

My appreciation to all the Scientists that I have worked with in the past especially Prof. Irene Evans (B.Sc Supervisor), Prof Fevzi Daldal (M.Sc Supervisor) and Dr Andreia Verissimo for teaching me lots of laboratory skills that and for prepping me to endure PhD journey.

Special thanks also to UMBI-UKM and MoE, in particular, Prof Rahman Jamal for giving me this opportunity and for awarding me the fellowship and sponsorship to pursue my PhD studies.

Finally, I would like to thank my Mum (Badariah) and Dad (Mohtar), for believing in me, for allowing me to leave the country for the third time (this time is 'closer' to home!) and for giving me freedom to do what I love. My humble gratitude to my parent-in-law: Azie and Jasni for allowing me to take your daughter for a very long honeymoon in the UK.

Abstract

Anterior gradient-2 (AGR2) is an endoplasmic reticulum (ER)-resident protein that belongs to a member of protein disulphide isomerase (PDI) superfamily. AGR2 initially emerged as a dominant effector of basic biological properties in vertebrates such as specifying forebrain integrity and limb generation. Subsequent studies in mammals implicated the role of AGR2 as a pro-metastatic protein essential to cancer progression and drug resistance, asthma and inflammatory bowel disease. AGR2 protein is mainly overexpressed in a number of human cancers and involved in pathways for ER stress, protein folding, transcription regulation, and exosome formation. Hence, AGR2 protein represents a clinically-relevant oncoprotein in tumour emergence and survival.

The aim of this thesis is to shed more light on the role of AGR2 in cancer development. AGR2 was previously shown that it could bind sequence-specifically to a linear peptide motif. In this study, hydrogen/deuterium exchange mass spectrometry was used to identify the dominant peptide-binding site on AGR2 by comparing the deuterium uptake between AGR2 and AGR2 with its ligand (linear peptide motif). The binding of the peptide was probed by making mutant series in the identified peptide 'docking site' region on AGR2. A consensus peptide-binding motif was then developed to identify potential cellular proteins that harbour this motif as potential AGR2 client proteins. Database mining using this consensus binding peptide demonstrated that transmembrane proteins were dominant class of proteins. Epithelial cell adhesion molecule (EpCAM), an oncogenic transmembrane protein, was chosen as putative AGR2 client proteins and their interaction was verified using both cell-free and cell-based assays. The AGR2 and EpCAM pathway dynamics were reconstituted and investigated in cells that do not endogenously express both proteins. Further, the expression of AGR2 and EpCAM were assessed in clinical tumour samples using immunohistochemistry. Proteomics screen using quantitative tandem mass tag (TMT) mass spectrometry on cells transiently overexpressing AGR2 were used to identify potential AGR2 client proteins and to find relevant dominant pathways affected by AGR2 signalling. Additionally, synthetic tools were devised to further dissect the function, regulation and 'druggability' of AGR2 protein. These tools include: i)

isolation of high-affinity AGR2-binding synthetic antibodies from a phage-scFv (single-chain variable fragment) library; ii) engineering synthetic mini-protein (synPRO) containing copies of wild-type and mutated AGR2 linear peptide motif and; iii) engineering synthetic membrane protein model that bind to AGR2.

In conclusion, the data presented hereby demonstrated a novel role of an ER-resident protein AGR2 which possess an intrinsic sequence-specific peptide binding for a subset of its client proteins and one function of this motif is to ensure proper maturation of client proteins to their final destination. Development of synthetic tools in this study can be further manipulated to disrupt AGR2 signalling and the fate of its binding proteins which in turn highlights a potentially 'druggable' stage in the oncogenic secretory pathway.

Lay Summary

AGR2 is an endoplasmic reticulum (ER)-resident protein that belongs to a member of protein disulphide isomerase (PDI) superfamily. AGR2 protein has emerged as a clinically-relevant protein as it is upregulated in various types of cancer and demonstrated tumour-promoting features.

The role of AGR2 is not fully understood, and it is the aim of this thesis to elucidate the role of AGR2 in cancer development. We focused on identifying novel AGR2 interacting proteins (termed interactome) as by looking at functional consequences of some of these interactions, we can deduce the role of AGR2. We used two approaches to identify AGR2 interacting proteins; 1) database mining of AGR2 sequence-specific peptide motif; 2) quantitative mass spectrometry analysis of cell models with a controlled expression of AGR2. We validated the key AGR2 client proteins using both *in-vitro* and cell-based assays.

Finally, we developed synthetic tools to further dissect the pro-oncogenic role of AGR2. These tools include: 1) developing AGR2 monoclonal antibodies in the format of the scFv fragment; 2) engineering synthetic mini-protein containing AGR2 sequence-specific peptide motif and; 3) engineering synthetic membrane protein that optimally binds to AGR2.

Table of Contents

List of Figures

List of Tables

Abbreviations

Ab	Antibody
AGR2	Anterior Gradient 2
AREG	Amphiregulin
ATP	Adenosine Triphosphate
bp	Base pair
BSA	Bovine serum albumin
CARD	Caspase activation and recruitment domains
Cas9	CRISPR associated protein 9
CCD	Coiled-coil domain
CHIP	C-terminus of Hsc70-interacting protein
CRISPR	Clustered regularly interspaced short palindromic repeats
DBD	DNA binding domain
DCD	Dermcidin
DMSO	Dimethyl sulfoxide
DMEM	Dulbecco's Modified Eagle Media
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DTT	Dithiothreitol
ECL	Enhanced chemiluminescence
ECM	Extracellular matrix
EGFR	Epidermal growth factor receptor
ELISA	Enzyme-linked immunosorbent assay
EpCAM	Epithelial cell adhesion molecule
ER	Endoplasmic reticulum
ERAD	ERassociated protein degradation
ERES	ER exit sites
ERGIC	ER-to-Golgi intermediate compartment
ERQC	ER quality control
EV	Empty vector
FACS	Fluorescence-activated cell sorting

FASP	Filter Aided Sample Preparation
FBS	Foetal bovine serum
FL	Full-length
gRNA	Guide RNA
GST	Glutathione S-transferase
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HPLC	High-performance liquid chromatography
HRP	Horseradish peroxidase
HSP	Heat shock protein
IF	Immunofluorescence
IP	Immunoprecipitation
IPTG	Isopropyl β-D-thiogalactopyranoside
kDa	Kilo Dalton
mAb	Monoclonal antibody
pAb	Polyclonal antibody
MDM2	Murine Double Minute 2
ml	Mililiter
mM	Milimolar
MOPS	3-(N-morpholino)propanesulfonic acid
mRNA	Messenger RNA
MS	Mass spectrometry
MUC	Mucin
OAC	Oesophageal adenocarcinoma
OD	Optical density
PBS	Phosphate-buffered saline
PBS-T	PBS Tween-20
PCR	Polymerase chain reaction
PDI	Protein disulphide isomerase
pfu	Plaque-forming unit
pH	Potential of hydrogen
PLA	Proximity ligation assay
PPI	Protein-protein interaction

RING	Really Interesting New Gene
RLU	Relative light unit
RNA	Ribonucleic acid
RPMI	Roswell Park Memorial Institute
RT	Room temperature
RT-PCR	Reverse Transcription PCR
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide pel electrophoresis
siRNA	Small interfering RNA
SP	Secretory pathway
SRP	Signal recognition particle
SWATH	Sequential window acquisition of all theoretical mass spectra
TEMED	Tetramethylethylenediamine
TMT	Tandem mass tag
Tris	Tris(hydroxymethyl)aminomethane
UPR	Unfolded protein response
WT	Wild type
v/v	Volume/volume
w/v	Weight/volume
Y2H	Yeast two-hybrid
µg	Microgram
µl	Microliter
µm	Micrometer

CHAPTER 1: Introduction

1.1. Cancer

1.1.1. Overview of cancer

Cancer refers to a class of diseases in which cells exhibit abnormal growth and can invade surrounding healthy tissues or spread to distant part of the body. Cancer is also known as a malignant tumour, where it can be differentiated with a benign tumour which refers to a tumour that does not invade or spread to another part of the body [1]. According to World Cancer Report 2014, published by the World Health Organization's International Agency for Research on Cancer[1]. Cancer is responsible for 8.2 million deaths which account for 13% of all deaths worldwide. This figure is expected to rise by about 70% over the next two decades. The world's most common cancer is lung cancer, making up 13% of the total number in 2012. The five most common type of cancer for men are lung, prostate, colorectal, stomach, and liver cancer while the top five most common type of cancer for women are breast, colorectal, and lung, cervix, and stomach cancer.

Cancer progresses from a single cell which transforms from normal into tumour cells through multi-stage processes, typically a progression from a pre-cancerous lesion to malignant tumours and involves different and variable activation and inactivation pathways. The transformation of a 'normal' cell to a cancerous cell requires a cascade of genetic alteration of genes modulating cell cycle, which are being silenced or overexpressed. Cancer can spread from its original site by local spread, lymphatic spread to regional lymph nodes or through the bloodstream to distant sites, known as metastasis. When cancer cells metastasize, they usually do so to preferential organs, depending on the type of cancer as hypothesised in the 'seed and soil' theory of cancer metastasis [2]. The process of cancer metastasis consists of a long series of sequential, interconnected steps and is dependent on both the intrinsic properties of the tumour cells and the responses of the host [3]. The symptoms of metastatic cancers depend on the location of the tumour and can include among others enlarged lymph

[1] http://www.who.int/cancer/en/

nodes, enlarged liver or enlarged spleen, pain or fracture of affected bones, as well as neurological symptoms.

There are many different kinds of cancer (more than 200 have already been classified) associated with almost every cell type, establishing cancer as a multistage and widely spread 'disease' [2]. Genetic abnormalities cause most cancers which can be due to the effect of carcinogens, UV, radiation, chemicals, tobacco smoke or viruses. Genetic abnormalities, including DNA mutations, also involve inheritance of certain cancer-promoting oncogenes or cancer-protecting tumour suppressor genes.

Classification of cancer is usually according to the organs or tissue from which the cells originate (primary tumour), as well as the specific type of cells that they originate from. Cancer diagnosis usually requires medical examinations such as blood tests, X-rays, endoscopy and computed tomography (CT) scans. This is followed by histological examination of a tissue biopsy or surgery specimen by a pathologist to determine the histological grade of the tumour along with other characteristics. This information helps the clinician to set prognosis of cancer patients. Prognosis is influenced by the type of cancer, as well as the stage, or extent of the disease. Additionally, the presence of specific molecular markers can also be useful in establishing a prognosis, as well as in determining individual treatments.

1.1.2. Hallmarks of cancer

The hallmarks of cancer, first defined in 2000 by Hanahan and Weinberg [4], provided a major foundation for the understanding of the complexity of this neoplastic disease (Figure 1.1). These traits that enable malignant growth include: (1) sustaining proliferative signalling, (2) evasion of programs that limit cell growth and (3) evade cell death, (4) unrestricted replicative potential, (5) induce angiogenesis and (6) trigger invasion and metastasis. An increasing research advance suggests that two additional emerging hallmarks of cancer are involved in the pathogenesis of some and perhaps all cancers [5]. These two emerging hallmarks involve (7) reprogramming energy metabolism (8) and circumventing the immune destruction. Additionally, two

[2] http://www.cancerresearchuk.org

consequential characteristics of neoplasia facilitate the acquisition of both core and emerging hallmarks (enabling characteristics) which include (9) genome instability and (10) tumour promoting inflammation (Figure 1.1).

1.1.3. Cancer targeted-therapy

Cancer can be treated by surgery, chemotherapy, radiotherapy, immunotherapy, monoclonal antibody therapy, or even combination of some methods. The choice of therapy depends on the location and grade of the tumour and the stage of the disease, as well as the general state of the patient. Surgery is the most effective and extensive approach in the treatment of primary tumour and affected lymph nodes because surgery operates by excising the tumours that are eventually killed. With the introduction of radiation therapy in the 1920s and chemotherapy after the 1940s, cancer surgery has become conservative. In contrast to surgery, chemotherapy and radiation therapy are only capable of killing a fraction of tumour cells by each treatment, and often a combination of all methods are necessary to remove cancer cells effectively.

Target-specific therapies have been developed for each emergence of hallmarks of cancer. One of the most utilised approaches towards developing an effective drug has been to target a very specific property or signalling pathway in the cancer cell that discriminates it from the normal cell. Targeted therapy approaches aim to inhibit molecular pathways that are crucial for tumour growth and progression; whereas, immunotherapy purposes to stimulate a host immune response that can obliterate the prolonged tumour growth. The combination of both immunotherapy and targeted therapy have been proposed to improve clinical outcomes [6] efficiently. Moreover, as the body of research advances and understanding of human cancer landscape progresses, treatments are refined for effective cancer therapy. There has been significant progress in the development of targeted therapy drugs that act specifically on molecular abnormalities in certain tumours, and which minimise damage to normal cells. Antibody therapy, for example, is targeted therapy strategy in which the therapeutic agent binds specifically to a protein on the surface of the cancer cells. This approach is possible due to human cancer cells express specific cell surface

antigens that are either overexpressed, mutated or selectively expressed compared with normal tissues [6]. Examples include the anti-HER2/neu antibody trastuzumab (Herceptin) used in breast cancer, and the anti-CD20 antibody rituximab, used in a variety of β-cell malignancies (Table 1.1)[7] [8]. Targeted therapy can also involve small molecules that act as "homing devices" which can target various phases in the transduction of the growth signal and its execution [7, 9]. Inhibitors are also the type of small molecule targeted therapy that has catalytic activity on mutated, overexpressed, or otherwise critical proteins within the cancer cell. Well-known examples include the tyrosine kinase inhibitors imatinib and gefitinib (Table 1.1).

Despite all the methodologies to treat cancer, the most deaths of cancer patients result from the progressive growth of metastases that are resistant to current therapies [10]. Metastatic cells can undergo an enormously high rate of spontaneous mutation in comparison to benign cells and hence, have a greater propensity to undergo rapid phenotypic modification that in turn become resistant to various therapeutic approaches [11].

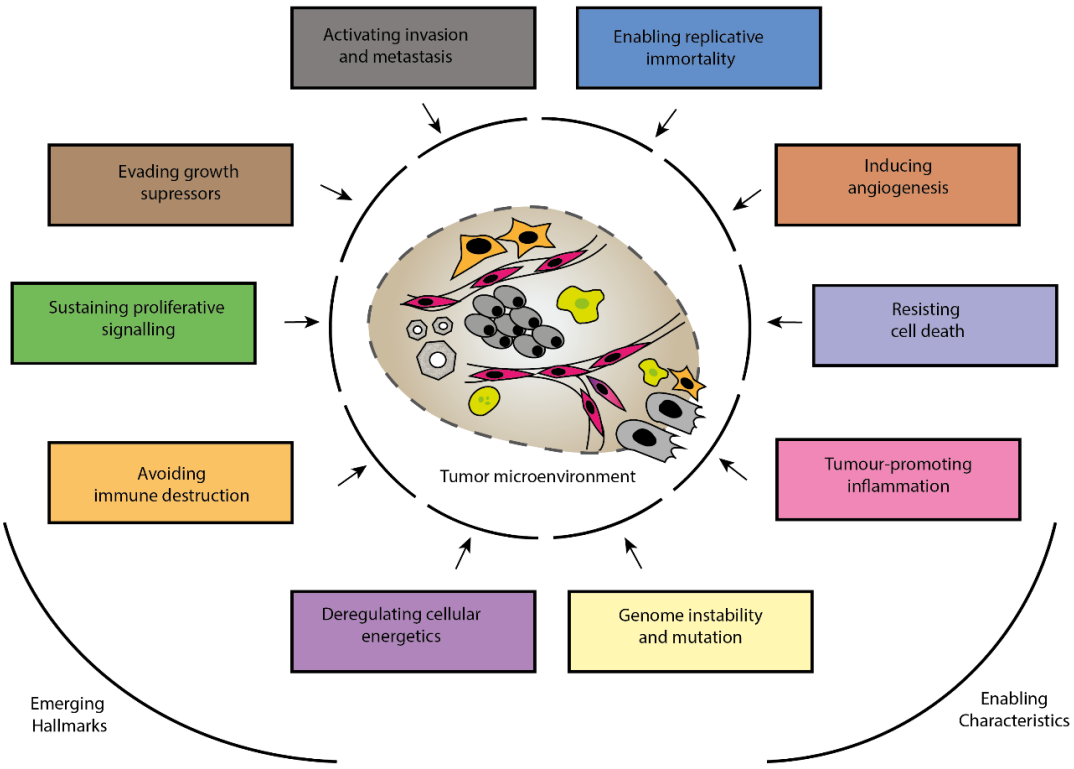


Figure 1.1 Hallmark capabilities of cancer. The six hallmarks of cancer originally proposed by Hanahan and Weinberg [4] acquired during the multistep progression of human tumours. The increasing form of research adds two additional emerging hallmarks [5]. Additionally, two enabling characteristics of cancer have been proposed that facilitate the acquisition of both core and emerging hallmarks [5] that altogether comprise a unifying principle for understanding the multiplicity of neoplastic disease.

THERAPEUTIC GROUP	DRUG	EFFECT ON TUMOUR
MONOCLONAL ANTIBODIES	Trastuzumab	Blocks growth signalling through HER2
	Bevacizumab	Neutralising antibody against VEGF: blocks angiogenesis
	Cetuximab	Neutralising antibody against EGFR: blocks growth signals
SMALL MOLECULES/INHIBITORS	Sunitinib	Blocks multiple tumour-associated tyrosine kinases, including VEGFR and PDGFR
	Imatinib	Blocks multiple tumour-associated tyrosine kinases, including ABL and KIT
	Vemurafenib	Blocks BRAF-V600E
	Temsirolimus, rapamycin and other mTOR inhibitors	Blocks mTOR pathway
	Bortezomib	Blocks 26S subunit of the proteasome
	JAK2 inhibitors	Block JAK2 signalling in tumour cells
	HSP90 inhibitors	Blocks HSP90, which increases unfolded protein-associated stress in tumour cells
	PI3K–AKT inhibitors	Decreases PI3K–AKT signalling in tumour cells
	Lenalidomide	Not well understood
	GSK3β inhibitors	Blocks GSK3β-mediated signalling of tumour cell growth
	IAP inhibitors	Sensitises tumour cells to apoptosis

Table 1.1. Overview of targeted therapy group for cancer treatment and their targets. (Reproduced from [7, 9])

1.2. Emerging strategy: targeting secretory pathway in cancer

1.2.1. Overview of the secretory pathway

The secretory pathway (SP) is a system to deliver proteins to the plasma membrane and specified membrane-bound organelles as well as to release material into extracellular space. Most of the cellular transmembrane proteins use this pathway to reach their final destination. Proteins destined for SP pass through membrane-bound compartments mediated by secretory vesicles that bud from one compartment and then fuse with the successive compartment. The SP consists of the rough endoplasmic reticulum (rough ER), ER exit sites (ERESs) the ER-to-Golgi intermediate compartment (ERGIC), the Golgi complex and secretory vesicles that transport proteins to final destination [12, 13].

Approximately one-third of the proteins in eukaryotes are targeted to the secretory pathway. The first compartment encountered by the SP substrates is the ER [14]. The ER contains molecular chaperones that orchestrate different maturation steps including protein folding, glycosylation, disulphide bond formation and catalysed co- and post-translational modifications ensuring that secretory proteins are properly folded, modified and assembled into multi-protein complexes in the ER. ER quality control (ERQC) governs these multi-steps processed inside the ER to ensure that the ER assembly line meets specific requirements [15]. ERQC comprises molecular chaperones that not only help proteins to fold but also assess the conformations of their substrates. If a protein has accomplished its native conformation, it might proceed to next subsequent step in the SP.

Most of the proteins for secretory pathway contains signal peptide at their N-terminus for targeting into ER and subsequent steps in the SP. This pathway is called the conventional secretory pathway (Figure 1.2). The conventional secretory pathway is the trafficking route that secretory proteins undertake when are transported from the ER to the Golgi apparatus network by successive fusion through the Golgi compartments, and subsequently to the plasma membrane via secretory vesicles or secretory granules. This process is mediated by many proteins, enzymes and co-factors. Similarly, the principal mechanism by which integral and transmembrane

proteins mature and arrive at their resident location is also through the secretory pathway. Nascent proteins that are being synthesized by ribosomes in the ER must be inserted into the ER membrane. Most of the proteins for secretory pathway contains signal peptide at their N-terminus. The N-terminal signal peptide (or leader sequence), is an ER-targeting sequence and hydrophobic, which is recognised by the signal recognition particle (SRP) complex in the ER membrane, which then targets the nascent polypeptide to an ER membrane component known as a translocon. The ER membrane maintains its integrity through the action of the BiP chaperone on the lumenal side, which associates with the translocon, moving away when the ribosome that is synthesising the nascent polypeptide associated with the cytoplasmic side of the translocon [16]. In the ER, the N-terminal signal sequence is cleaved off, and proteins are folded and assembled by molecular chaperones whilst preventing unwanted protein aggregation due to exposure of hydrophobic patches to the aqueous environment [17].This is to ensure that the hydrophobic portions of the nascent polypeptide are shielded from the aqueous environment, whilst allowing the hydrophilic portions destined for the cytoplasm or the extracellular environment to pass through to the ER lumen or cytoplasm. The newly synthesised proteins then exit the ER at dedicated membrane domains, ERESs which are ribosome-free regions of the rough ER. The ERESs house the cargo-containing coat protein complex II (COPII)-coated vesicles form along with the association of plus end microtubules [18]. After formation, COPII vesicles can fuse homotypically to generate pre-Golgi intermediates [19]. Alternatively, COPII vesicles may heterotypically fuse with the pre-formed ERGIC [20]. Subsequently, the newly synthesised proteins reach the Golgi apparatus where they are further modified by glycosylation and/or processing, sorted and dispatched towards their correct final destinations. The Golgi apparatus comprises a collection of fused, flattened membrane-bound disks that are called cisternae. Golgi stacks are made up of a series of compartments consisting of two main networks: a reticular membrane network on the entry side of the Golgi that builds the cis-Golgi network (CGN) and the trans-Golgi network (TGN) on its exit side [21]. At the TGN, proteins destined to be secreted are sorted in secretory vesicles (SVs) or immature secretory granules (ISGs). SVs are constitutively delivered toward the PM, whereas ISGs accumulate in the cytoplasm. Upon the arrival of specific stimuli, ISGs form mature secretory

granules (MSGs) that are transported to the PM.Proteins can also be returned to the ER through COPI-coated vesicles via retrograde transport involving components in the Golgi and back to the ER [22]. COPI-coated vesicles also partake in the anterograde movement of certain components which occurring in ERGIC region [23]. One significant function of the COPI-coated vesicle is to facilitate the retrieval of escaped luminal proteins containing K/HDEL retrieval signals that are recognised by the K/HDEL receptor as well as other machinery required for optimal anterograde transport. The transport of secretory cargo between the ER and the Golgi occur bi-directionally involving COPII and COPI protein complexes. Trafficking of the proteins within the SP also involves a series of fusion events that occur between secretory vesicles with their target membrane-bound compartments such as a lysosome and plasma membrane. These fusions are catalysed by SNAREs (soluble N-ethylmaleimide-sensitive fusion protein (NSF) accessory protein (SNAP) receptors) which contain long regions that form helical structures[21]. Membrane fusion is mediated by the formation of tight trans-SNARE complexes that are formed by α-helical bundles of the cytoplasmic SNARE domains (v-SNAREs) and target SNAREs (tSNAREs) that interact when brought into proximity and appear to zipper up and eventually results in membrane fusion [21, 24].

Some signal-peptide-containing proteins have been shown not to follow this conventional secretory pathway. These processes are known as unconventional protein secretion (Figure 1.2) [25]. There are some evidences that some proteins do not seem to involve COPII vesicles. For example, in the case of yeast protein Heat-shock protein 150 (Hsp150), the protein was seen transported to the plasma membrane upon exit the ER without COPII coat proteins Sec24 and Sec13 [26]. Some proteins also were shown to be inserted to ER but bypass the Golgi apparatus. For example, the cystic fibrosis transmembrane conductance regulator (CFTR) which reaches the plasma membrane by direct transport from the ER by COPII carriers [27]. However, the exact mechanisms and protein components of unconventional protein secretion are not known and are beginning to emerge [28].

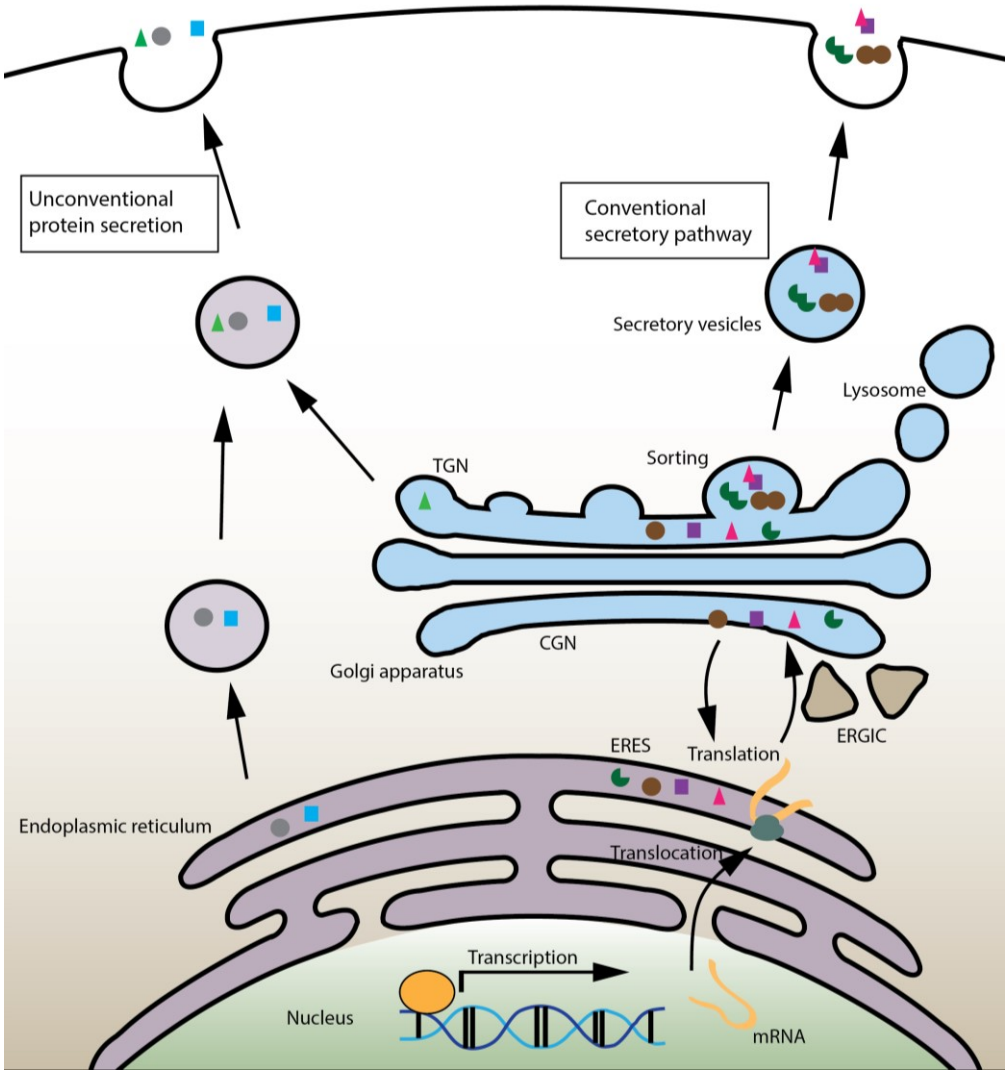


Figure 1.2 Overview of the secretory pathway. Most of the proteins enter secretory pathway by possessing signal peptide at their N-terminus for targeting into ER followed by the ER-Golgi system and subsequent steps in SP. This is called the conventional secretory pathway. However, there are some signal-peptide-containing proteins not to follow this conventional secretory pathway. This is known as unconventional protein secretion which can occur by either bypassing the association of COPII vesicles during ER exit or bypassing Golgi apparatus (transport from the ER to the plasma membrane).

1.2.2. Unfolded protein response

The SP involves multistep and active processes that are closely regulated and controlled to maintain ER protein homoeostasis or proteostasis [29, 30]. The ER particularly plays an important role in the folding and maturation of newly synthesised secretory and transmembrane proteins. The ER lumen governs a unique environment to establish a balance between the ER protein load and the capacity to handle this load in order to ensure proper protein-folding requests. This ER homeostasis can be perturbed by physiological as well as multiple environmental and cellular signals such as high protein demand, viral infections, environmental toxins, inflammatory cytokines, and mutant protein expression resulting to an accumulation of misfolded and unfolded proteins in the ER lumen, a condition termed as ER stress. It is known that under ER stress, cells activate a series of complementary adaptive mechanisms to cope with increased demands of protein folding in the ER. This adaptive system is called the unfolded protein response (UPR) which is a highly conserved signal transduction pathway.

The UPR signals protein-folding status in the ER lumen to the nucleus and cytosol to buffer fluctuations in unfolded protein load [31]. When unfolded proteins accumulate in the ER, resident chaperones become occupied, releasing transmembrane ER proteins involved in inducing the UPR. These proteins engaged ER membranes, with their N-terminus in the lumen of the ER and their C-terminus in the cytosol, providing a bridge that connects these 2 compartments. Generally, in unstressed condition, the ER chaperone Grp78 (BiP) binds to the N-termini of these transmembrane ER proteins, preventing their aggregation. Upon accumulation of misfolded proteins, Grp78 dissociates from these transmembrane ER proteins, allowing aggregation of these sensors, and as a result activating the UPR. The UPR adaptive response includes upregulation of molecular chaperones and protein processing enzymes to increase folding and handling efficiency, translational attenuation to reduce ER workload and prevent further accumulation of unfolded proteins, and an increase in ER-associated protein degradation (ERAD) and autophagy components to promote clearance of unwanted proteins. The process of ER stress in cells can be irreversible [32], which can lead the cells being eliminated by apoptosis.

Cell damage due to chronic ER stress is widely common to a wide range of dominant human diseases, including cancer, neurodegeneration, stroke, pulmonary fibrosis, viral infections, inflammation, metabolic disorders, diabetes, and heart disease [33].

There are three main players in mammalian UPR which comprise the sensors IRE1α (endoribonuclease inositol-requiring enzyme 1-alpha), PERK (protein kinase RNA-like endoplasmic reticulum kinase) and ATF6α (activating transcription factor 6) [34].These three arms of UPR activate stress signals from the ER lumen to the rest downstream signalling that allow cells to cope with protein workload and misfolding by provisionally reducing *de novo* protein synthesis and improving the folding and clearance capacity of the ER [35].

The first UPR sensor IRE1α is an ER-resident type 1 transmembrane protein. It has two enzymatic activities in its cytosolic domain: a serine/threonine kinase and an endoribonuclease activity. Upon activation, IRE1α dimerizes to form homodimer followed by an auto transphosphorylation, which triggers an unconventional RNA splicing activity, which processes the mRNA transcription factor called unspliced X-box-binding protein 1 (XBP1u) to produce an active transcription factor, spliced XBP1 (XBP1s). XBP1s controls the transcription of genes encoding proteins involved in protein folding, ER-associated degradation (ERAD), protein quality control and phospholipid synthesis. IRE1α also degrades the mRNA of several secreted proteins, through a process called regulated IRE1-dependent decay (RIDD). Besides, IRE1α can signal through the recruitment of TRAF2 leading the activation of the ASK1–JNK cascade.

PERK is a transmembrane protein with a specific kinase activity in its cytosolic domain. Its main substrate is the translation initiation factor eIF2α. Phosphorylation of eIF2α results in a decrease in translation as well as a preferential translation of key proteins such as CHOP and GADD34, two factors directly involved in the cellular decisions of life or death. PERK can also phosphorylate nuclear factor erythroid 2-related factor 2 (NRF2), a transcription factor involved in redox metabolism.

Finally, ATF6 is a transmembrane protein and an ER transcription factor. Under normal ER condition, ATF6 has a basic Leu zipper (bZIP) transcription factor in its cytosolic domain and localises in the ER. When ER is under stress condition, ATF6 is exported to the Golgi apparatus, through interaction with the coat protein II (COPII) complex, and finally cleaved from its membrane by the proteases S1P and S2P, releasing its cytosolic domain fragment (ATF6f). ATF6f pedals the upregulation of genes encoding ERAD components and also XBP1.

Together, integration signals from these three UPR arms determine cell survival by the regulation of distinct subsets of gene transcription and translation reorientation toward recovery of ER homoeostasis. One of the first responses from UPR is the downregulation of translation to reduce entry of newly synthesised proteins in the ER. Among other cellular processes include the activation of amino acid metabolism and upregulation of antioxidant capacity, increased clearance of misfolded proteins, and increase expression of ER chaperones/ER quality control proteins involved in protein folding. However, if these adaptive actions exceed ER protein folding capacity, ER stress shifts towards the terminal UPR response which is apoptosis [36].

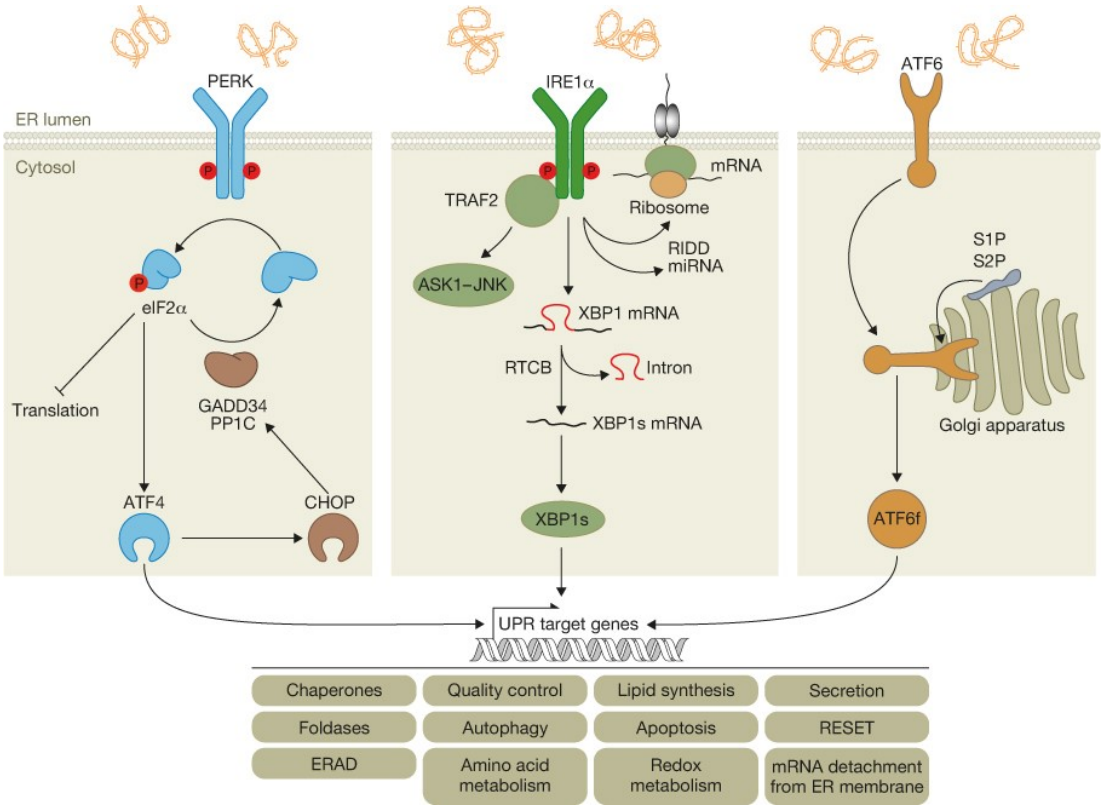


Figure 1.3 The three arms of UPR. (From [34]) All three of UPR arms (PERK, IRE1α, ATF6) are present in the ER membrane that initiates signalling events to maintain ER protein homoeostasis or ER proteostasis under ER stress. PERK phosphorylates eIF2α, which results in a decrease in translation as well as a preferential translation of key proteins such as CHOP and GADD34, two factors directly involved in the cellular decisions of life or death or concurrently increases the expression of the transcription factor ATF4. IRE1α form homodimer upon activation and signals through (i) the recruitment of TRAF2 leading to the activation of the ASK1–JNK cascade and (ii) through its RNase activity by splicing the XBP1 mRNA or the degradation of RNAs (RIDD activity), regulating gene expression at transcriptional and post-transcriptional levels. Finally, ATF6 normally localised at the ER in unstressed cells is exported from the ER to the Golgi complex upon activation through interaction with the coat protein II (COPII) complex, where it is cleaved by the proteases S1P and S2P, releasing its cytosolic domain releasing its cytosolic domain fragment (ATF6f). Depending on the severity and duration of ER stress, the signalling branches of the UPR can activate adaptive determine cell survival toward recovery or induce apoptotic cell death.

1.2.3. Cancer cell secretory pathway is addicted to protein secretion

Cancers development are triggered by different hallmarks of cancer [5]. Cancer cells grow under oncogenic stress caused by hypoxia, nutrient deprivation, DNA damage, metabolic stress, and oxidative stress, leading to UPR as an adaptation strategy. However, most normal cells are not subjected to stress and the UPR pathways remain inactive in these cells. ER resident proteins display altered expression patterns in cancer. For example the ER chaperone protein Grp78 (BiP) is one of the most active components of cancer cells and is overexpressed in different kinds of cancers [37]. Most cancers have to cope with the secretory crisis that demands secretory proteins to be secreted through SP. In recent years, an important link between ER carcinogenesis and tumour development has been established [38]. ER is the first compartment of the secretory pathway which orchestrate secretion and the cell microenvironment. In transformed cells like cancer, the SP is exposed to a strong environmental pressure such as hypoxia, oxidative stress or chemotherapies, aneuploidy, an increase in metabolic demand, and high proliferation rates which together contribute to tumorigenesis. This increased demand in tumours leads to the accumulation of misfolded proteins in the secretory pathway and more particularly in the ER. The requirement of a robust ER and secretory machinery are necessary for tumour cells, and the cellular challenges can trigger disturbance of ER homoeostasis and consequently causing UPR activation at the basal level or constitutively active [39]. The increase in secretion demand is also dependent on a substantial requirement of energy and amino acid supply which can be highly costly in cancer environment [39].

There are several consequences of altered SP in transformed cells as described by Dejeans *et al.* (Figure 1.4) [40]. First, it can lead to microenvironment landscape destabilisation by remodelling of the extracellular matrix (ECM) through changes in ECM components abundance or matrix metalloproteinases (MMPs) expression which linked to cancer dissemination and invasion. Second, there is a modulation of signalling proteins or contact/adhesion protein abundance which can trigger cancer cell proliferation, migration or tumour angiogenesis. Third, it can lead to an alteration of the immune response by expanding the proinflammatory response of tumour cells.

Finally, it can remodel tumour immunogenicity by stimulating plasma membrane exposure of major ER -resident chaperones such as calreticulin (CRT).

The activation of UPR signalling can attract both pro-survival and pro-apoptotic cellular settings. Therefore, therapeutic intervention targeting the secretory pathway is a fascinating strategy to target cancer cells (Figure 1.5) [39, 41]. Cancer drugs or chemical probes can be designed for example by; (i) reducing the amount of protein load entering the secretory pathway or neutralizing the increased demand for secretory proteins mediated by hyperploidy or oncogene activation (ii) controlling the secretion of specific set of proteins, such as receptor tyrosine kinases, pro-inflammatory messenger, by modulating protein processing at post-translational level (iii) targeting ER chaperones, such as BiP or GRP94 that show chemoresistance characteristic, (iv) upregulating macroautophagy or enhance the efficiency of the ER-associated protein degradation (ERAD) pathway (v) and modulating the UPR (inhibitors and activator) to sensitize cancer cells to ER stress and cancer. Screening studies for example have identified several compounds that can inhibit key mediators of ER stress signalling [34]. For examples, 4μ8C, MKC-3946 and STF-083010 compounds were identified by high-throughput screenings that can target IRE1α. These compounds are in pre-clinical trials which can interact IRE1α via one of two sites: the catalytic core of the RNase domain or the ATP-binding pocket of the kinase domain. Another example of IRE1α inhibitors are Sunitinib and APY29. Both molecules bind to the hinge region of the ATP-binding pocket and stabilize an active kinase domain conformation. Sunitinib and APY29 block IRE1α activity but can allosterically activate the IRE1α RNase domain *in-vitro* [42]. Sunitinib (also called Sutent) was approved by the United States Food and Drug Administration (FDA) for targeted therapy of renal cell carcinoma (RCC) kidney, gastrointestinal stromal tumours and pancreatic neuroendocrine tumours (GISTs). Screening studies also has identified compound that can inhibit PERK. One example is GSK2606414, a small molecule that inhibits PERK phosphorylation. While still in preclinical trial, GSK2606414 displayed a potent antitumor and antiangiogenic activities in a xenograft model of pancreatic cancer [43].

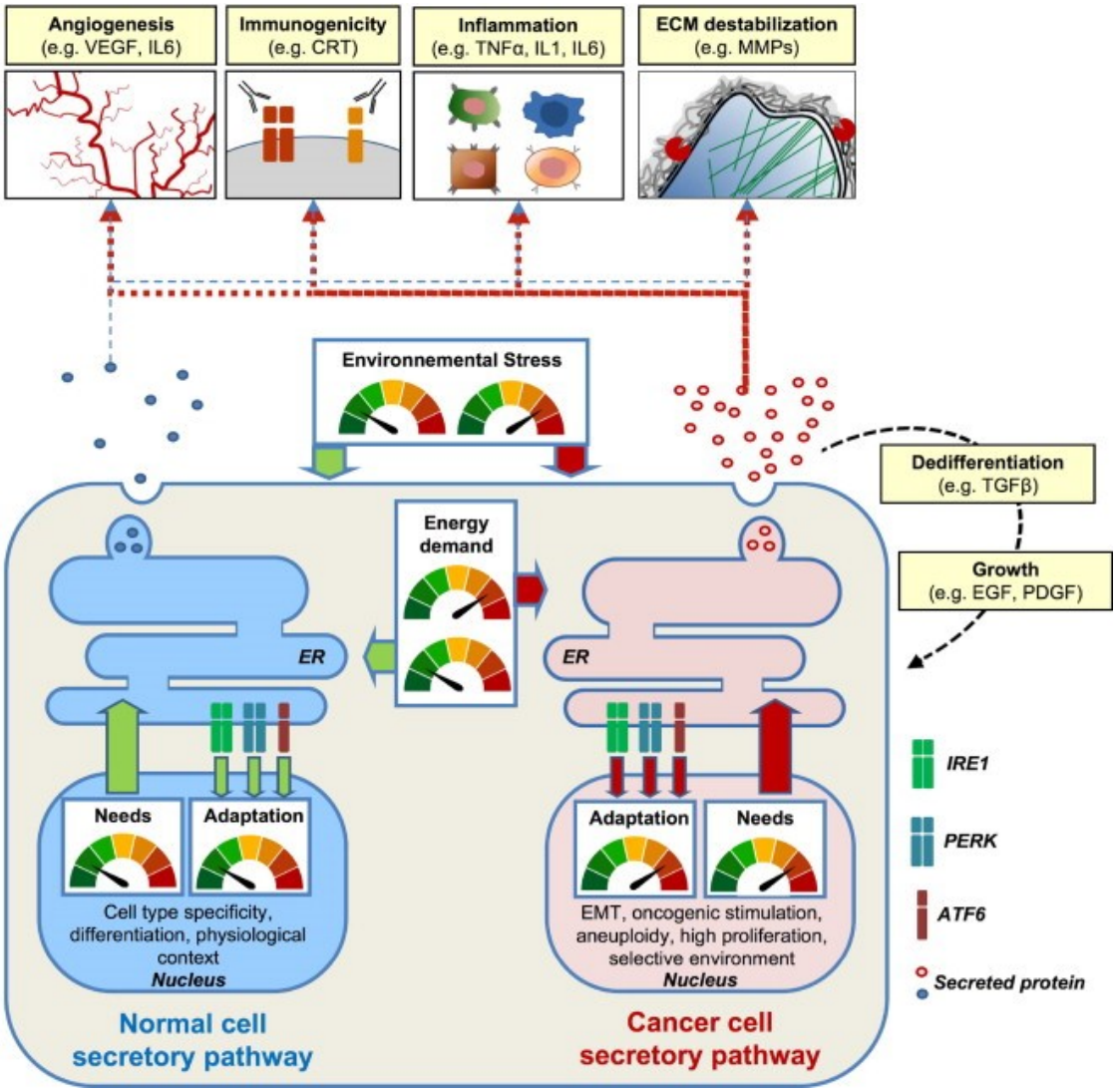


Figure 1.4 Cancer cell secretory pathway addiction to secrete. (From [40]) Protein secretion is dependent on both intrinsic (aneuploidy, increase in metabolic demand, and high proliferation rates) and extrinsic factors (environmental stress such as hypoxia or oxidative stress). In cancer cells, these factors are relatively high which leads to the accumulation of misfolded proteins in the secretory pathway and more particularly in the ER. There are several consequences of altered SP which include destabilisation of ECM, modulation of signalling proteins or contact/adhesion protein abundance, alteration of the immune response and remodelling of tumour immunogenicity.

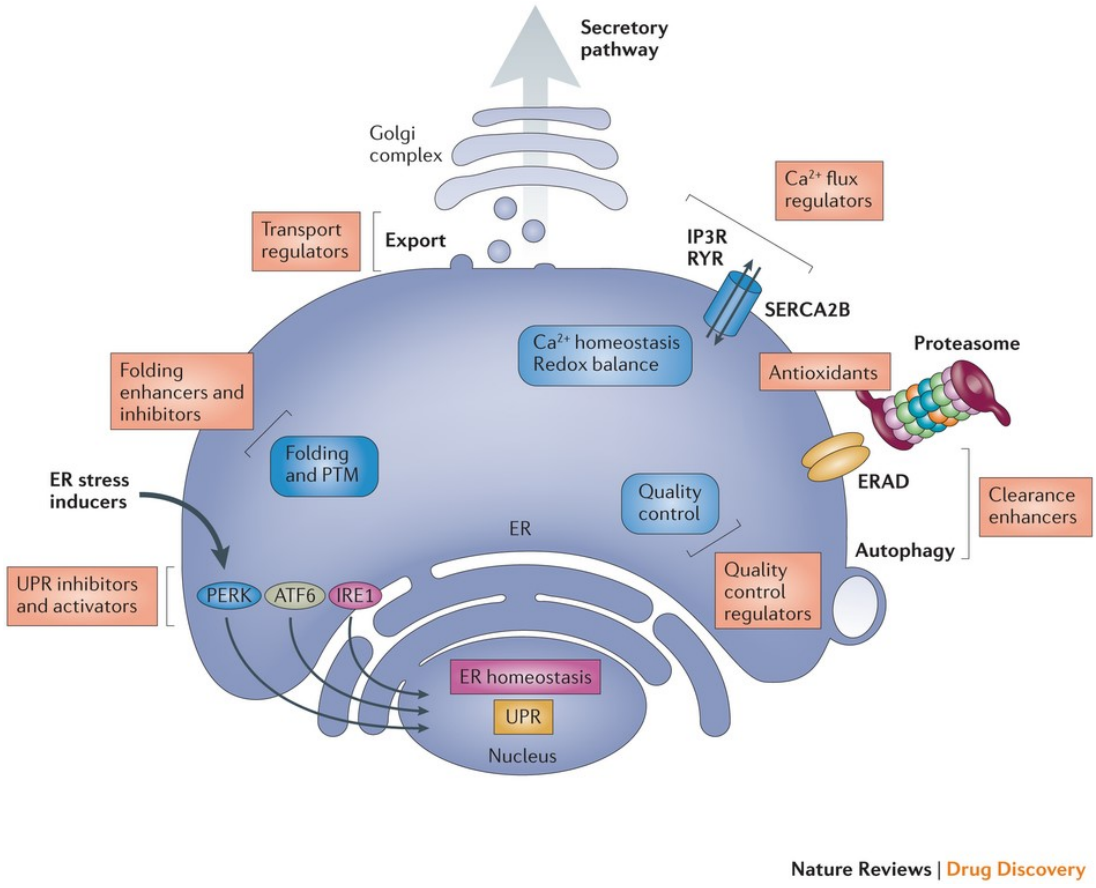


Figure 1.5 Targeting the secretory pathway in cancer. (From [41]) Therapeutic interventions are possible at many time points in the secretory pathway of cancer. Cancer drugs can be designed for example by; (i) reducing the amount of protein load entering the secretory pathway, (ii) modulating the secretion of select proteins, such as receptor tyrosine kinases, growth factors, or cytokines, (iii) targeting ER chaperones, such as BiP or GRP94 that show chemoresistance characteristic, (iv) upregulating macroautophagy or enhance the efficiency of the ER-associated protein degradation (ERAD) pathway (v) modulating the UPR (inhibitors and activator) to sensitize cancer cells to ER stress and cancer.

1.2.4. ERAD and interplay with autophagy during ER Stress and UPR

Despite the stringent requirements of ERQC, there is evidence to suggest that at least a third of all proteins entering the ER fail to satisfy the quality-control mechanisms that ensure proper folding, and for some proteins, the success rate is much lower[44]. Protein folding itself is prone to error because of the differences in protein folding and protein aggregation free energy landscapes[45]. Moreover, spontaneous errors during transcription and translation, genetic mutations, toxic compounds and cellular stresses, can contribute to folding efficiency. ERQC ensures that proteins are properly folded and ensure correct protein conformations. If folding is delayed or the specific protein conformations are not met, the proteins are either subjected to additional folding steps or selected for a process called ER-associated degradation (ERAD)[46]. As discussed above, ER stress for instance can lead to the build-up of unfolded, misfolded or damaged proteins that exceeds the capacity of chaperones available in the ER to correctly fold them. ERAD is a stringent quality-control system that involves in the clearance of these improperly folded proteins . During ERAD, molecular chaperones and associated factors recognise and target substrates for retrotranslocation and ubiquitination of misfolded proteins in the ER for proteasomal degradation in the cytosol to the cytoplasm for subsequent degradation by ubiquitin–proteasome machinery [47]. The ERAD system encompasses the transmembrane ubiquitin E3 ligase proteins that connect the substrate recognition machinery in the ER lumen and the ubiquitin-proteasome system in the cytosol.

ERAD is a regulated process which is important for maintaining ER homoeostasis. Several studies showed that there is an interplay between ERAD and other pathways[47]. The transcription of a certain subclass factors required for ERAD was shown to be induced by the UPR[48, 49], which is activated through the directed interaction of misfolded proteins with a UPR sensor IRE1 in the ER or through BiP which is titrated away to bind to misfolded substrates.

Autophagy is another pathway that reduces ER stress [47]. Substrates that accumulate over time are prone to aggregation which leads to ERAD inefficiency to clear this aggregation. This is where autophagy comes into play to reduce the

substrates aggregation by autophagy-mediated destruction. During this process, portions of the ER, along with proteins and protein aggregates, are surrounded in double-membrane structures called autophagosomes which will be then delivered to the lysosome or vacuole for degradation [50]. Similarly to the UPR, autophagy can result in either cell survival or cell death (167,168). The mechanisms by which the UPR induces autophagy are poorly understood, but possibly involve signalling through PERK–eukaryotic translation initiation factor 2α (eIF2α) and IRE1α169. The is an interplay between ERAD machinery with autophagy. Proteasomes which involves in ERAD machinery are multi-protein complexes made up of an inner 20S cylindrical shaped core of subunits that have the proteolytic activity and a 19S regulatory cap of subunits on each end of the core which recognize ubiquitin-conjugated proteins and provide ATPase activity [36]. Proteasomes and autophagy are co-dependent system that partaken in the digestion of soluble proteins and releasing peptides into the cytoplasm and nucleus to be digested by peptidases, or to be transferred to the ER where they are bound by human leukocyte antigen (HLA) proteins to the surface of the cell [25, 64]. The proteasome degrades type I (short-lived) and type II (misfolded/dysfunctional) proteins, whereas autophagy degrades type II and type III (long-lived) proteins [14]. This suggests that there is a backup mechanism for ERAD for clearance of misfolded proteins especially those related to substrate aggregations. The cellular ubiquitination system specifies whether a protein will be transferred to the proteasome or an autophagosome for recycling. Proteins with attached polyubiquitin chains that are branched on Lys48 have a more closed conformation and are targeted for proteasomal degradation, while proteins with single ubiquitin moieties or polyubiquitin chains that are branched on other Lys residues are targeted to the autophagosomes as described above [98, 121]. Some proteins can be digested by both the proteasome and autophagic vesicles [64].

1.3. Anterior gradient-2

1.3.1. Discovery of AGR2 and its key biological description

AGR2 belongs to the wider family of anterior gradient proteins including the ancestral member XAG-2. The anterior gradient XAG-1, XAG-2 and XAG-3, were first discovered in *Xenopus laevis* and named accordingly based on their specific expression patterns during early development [51]. XAG-2 was further investigated in a study to understand the patterning of anterior-posterior pattern formation in the neural plate. XAG-2 was described as a secreted proteins that are expressed in the anterior region of dorsal ectoderm from late gastrula stages onwards [52]. Gain-of-function analysis showed that XAG-2 function in specifying the fate of the dorso-anterior ectoderm that forms the mucus secreting cement gland and in inducing the forebrain's fate. Another *XAg-2* homologous gene, namely *XAgr-2*, was identified more recently in Xenopus embryos was also expressed in cement gland. Also, its expression was also found in the otic vesicles, specifically in the ventromedial region that later develops to the cochlea, and notochord, especially in the tailbud and neurula stages [53].

A rather late key insight of AGR2 function came from the studies of limb regeneration in the salamander. *Notophthalmus viridescens* (newt) homologue of *agr2* (nAG) was found as candidate genes induced during limb regeneration following amputation [54]. nAG was discovered as a secreted ligand for the newt GPI-anchored three-finger protein (TFP) Prod1 using yeast-2-hybrid studies [54], where Prod1 is membrane receptor which critical in determining the proximodistal identity of regenerating limb [55]. It was later described that nAG may interact with Prod1 through the activation of ERK1/2 MAP kinase signalling in blastema cells [56]. After newt amputation, nAG sequentially expressed in the Schwann cells of the distal nerve sheath and then in glands in the wound epidermis followed by release to the distal sheath and induces the formation of glands in the wound epidermis [54]. Further, when nAG is expressed via electroporation into denervated, and amputated limb was able to rescue denervated blastema and allow it to regenerate in the absence of the nerve. This indicates that nAG, secreted by nerve cells of the regenerating limb, specifically acts to differentiate limb shape and structure [54]. Surprisingly, no apparent homologs of

Prod1 were found in other vertebrates which lead to the question of whether nAG acts as a major regulator of regeneration in other vertebrates. The role of anterior gradient proteins in limb regeneration has since been broadened to other amphibians (namely *Xenopus laevis* tadpoles) and has been described as having a universal role in body appendage regeneration in amphibians [57]. Interestingly, human AGR2 (hAGR2) has been reported to interact with another GPI-anchored protein, metastasis protein C4.4 (now known as LYPD3), which shares some structural similarity with Prod-1 inferring to their cysteine-rich domains and Ly6-type domains [58]. This suggests that there might be a preserved, functional interaction pathway of protein complexes.

Another key insight of biological description of AGR2 functions came from the studies of mice. In situ hybridization revealed that mouse AGR2 (Gob4/mAGR2) is localised in the goblet cells of the intestine. Goblet cells, hence the name, are goblet-like shape differentiated epithelial cells located in the lining of the small intestine. Northern blot analysis of showed mAGR2 expression in the small intestines, colon and stomach of the adult mouse [59]. Later study also described AGR2 expression in the goblet cells of the intestine in mice [60]. The amino acid sequence of mAGR2 is similar to *Xenopus laevis* (46% homology and both contain a putative cleavable leader sequence required for targeting proteins to the secretory pathway)[59]. This finding gives a hint that that AGR2 might be involved in the secretory pathway. Additionally, both mice and Xenopus study showed the involvement of AGR2 in cement gland which consists of mucus-secreting cells. The cement gland, as suggested by the name, forms a cone of columnar epithelium that mediates the attachment to a solid support, which allows the Xenopus embryo to attach before it can swim [51]. This ‘adhesion’ inducing property of AGR2 predicts its potential role in migration and adhesion in mammalian cancer metastasis [61].

Studies using transgenic mice have begun to elucidate the role of mAGR2 in homoeostasis of the intestine. In normal intestine, AGR2 is found in several cell types of both proliferating and differentiated intestinal cells of secretory lineage which include mature goblet, Paneth and enteroendocrine cells goblet cell as well as goblet cell progenitors and MSI1-positive intestinal stem cells/early progenitors [62]. A null

mutation in the *Agr2* gene (*Agr2* -/-) in mice results in the loss of morphologically normal goblet cells and decreased intestinal mucus [60]. The null mutation of AGR2 also revealed the association of AGR2 with intestinal mucin MUC2, a large, cysteine-rich glycoprotein that forms the protective mucus gel lining the intestine, where there is no MUC2 synthesis detected [60, 63]. AGR2 co-immunoprecipitates with MUC2, a large, cysteine-rich glycoprotein that forms the protective mucus gel lining the intestine, and that this association depends on a conserved cysteine residue in the AGR2 thioredoxin-like domain [60]. The null mutation of AGR2 revealed the association of AGR2 with intestinal mucin MUC2, where there is no MUC2 synthesis detected [60, 63], suggesting that AGR2 protein remodels the cell to adapt to a demand for secretory protein synthesis. Mice lacking AGR2 were viable but were highly susceptible to colitis. Additional work by Zhao *et al.* using transgenic mice describes additional abnormalities causing symptoms in the small intestine of AGR2 knockout mice [63]. The Paneth cells morphology was significantly altered and showed severe intestinal inflammation consistent with the role of Paneth cells in the inflammatory response [64, 65]. Additionally, the Paneth cells display abnormal localisation, with differentiation away from the crypt base into the upper crypt and villi of the small intestine. Furthermore, loss of AGR2 expression leads to disrupted enterocyte homoeostasis, reduced proliferation, increased apoptosis and reducing villi [63]. Interestingly, cell culture experiments showed that ER stress induces AGR2 expression and that siRNA knockdown of AGR2 elevates ER stress response [63].

Altogether, these studies present fundamental paradigms, indicating the role of AGR2 and its homologues can function in epithelial cells as dominant endoplasmic reticulum resident molecular chaperone that can mediate proper protein maturation and maintain secretory landscape even in the presence of stress. AGR2 property in maintaining the epithelial barrier is further indicative of how it can function in diseases like cancer.

1.3.2. AGR2 and related family members

AGR2 belongs to the wider family of the anterior gradient (AGR) proteins which include AGR1 and AGR3. Human AGR2 and AGR3 proteins affiliation to the

PDI family were based on the initial phylogenetic analysis performed by Persson *et al.* where it was shown that both proteins have high homology to ERp18/19 protein, which now also denoted as AGR1 [66]. Persson *et al.* used BLASTP and BLASTN searches using the different PDIs/ERps from Homo sapiens to search for individual thioredoxin domains have evolved among the PDI/ERp family members [66].

Since there are more homologous *Agr* genes found in vertebrates, there is substantial confusion exists in the literature concerning the terminology and orthologous relationships between anterior gradient genes in different species. More recently, Ivanova *et al.* combined bioinformatics tools and expression analysis to gain more insight into the relationship [57]. Using a source from Ensembl and Genebank, they screened the sequences homologous to known anterior gradient genes and analysed local synteny of Agrs-containing loci in 17 amniotic and 6 anamniotic species while keeping the integrity of the orthologous relationships between members of Agr family in different species. They concluded that there are three subfamilies of *Agrs; Ag1, Agr2* and *Agr3* in which exist two syntenic fragments, *Ag1* and *Agr2/Agr3*. Interestingly genes of TLP19 subfamily which encodes non-secreted PDI are the nearest homologs (Figure 1.6A). Importantly, *Ag1* genes present in fishes and amphibians but lost in amniote ancestors (reptiles, birds, and mammals) which suggest a decreased ability to regenerate body appendages in higher vertebrates (Figure 1.6B). Using *Xenopus laevis* tadpoles as a model, amputation of the animals' hind limb buds and tails causes a substantial increase in expression which represents the two revealed syngenic groups of *Agr*: *Xag2*, *Xagr2a* and *Xagr3a* [57]. This finding was further complement with their *in-situ* hybridisation with *Xag2* and *Xagr2a* probes and by the *in-vivo* monitoring of regeneration in transgenic tadpoles expressing EGFP under the control of the *Xag2* promoter. Together with the data obtained in salamanders, indicate that *Agrs* have a universal role in body appendage regeneration in amphibians.

In addition, in a more recent study, Ivanova *et al.* the biological role of Ag1 [67]. It has been postulated that higher vertebrates have only *Agr2* and *Agr3*, but lack *Ag1* that is apparently lost from their ancestors which explain why higher vertebrates have a low ability to regenerate the body appendages. Therefore the loss of *Ag1* in

evolution could be an important event that led to a decline of the regenerative capacity in higher vertebrates. Using the dish *Dario rerio,* a group of lower vertebrates as a model, Ivanova *et al.* demonstrated that Dario Ag1 (*DAg1*) expression, like the expression of its *Xenopus* ortholog *Xag1*, increases after amputation of the body appendages which in this case the fins [67]. In concordance with this result, downregulating *DAg1* using Vivo-morpholino antisense oligonucleotides, resulted in an impedance of the fin regeneration [67]. This suggests that the loss of *Ag1* in higher vertebrates' ancestors could be a cause of the reduction of regenerative capacity in their modern descendants.

Another phylogenetic analyses through database mining of both transcriptomic and genomic databases have identified a novel group of Caudata (salamander) AG proteins, which were named AG4 [68]. The phylogenetic analyses suggested that AG proteins are vertebrate-specific and cluster in four groups each of which has a different distribution of vertebrate classes and has thioredoxin active site/C-terminus profile. AG4 sequences are slightly more similar to mammalian AGR2 than nAG (e.g. 75% vs. 69% identity between human AGR2 and *Notophthalmus* AG4 and nAG, respectively). The expression and secretion of the newly found gene AG4 were assessed using Western blot. AG4 along with constructs expressing wild-type nAG, or human AGR2 were transfected into cancer cells, and the media were collected after three days, concentrated, and analysed by western blotting in parallel with the cell lysates. The results showed that secreted AG4 was approximately 20-fold lower than that for nAG, while that for AGR2 was nearly 6-fold lower suggesting a significant difference in secretion between newt AG4 and nAG[68]. The author suggested that the resilt may be due to the co-existence of multiple AG proteins that might have moderately overlapping functions and each AG protein subfamilies have novel activities through evolution by modulating the cellular localisation and/or the degree of ER retention.

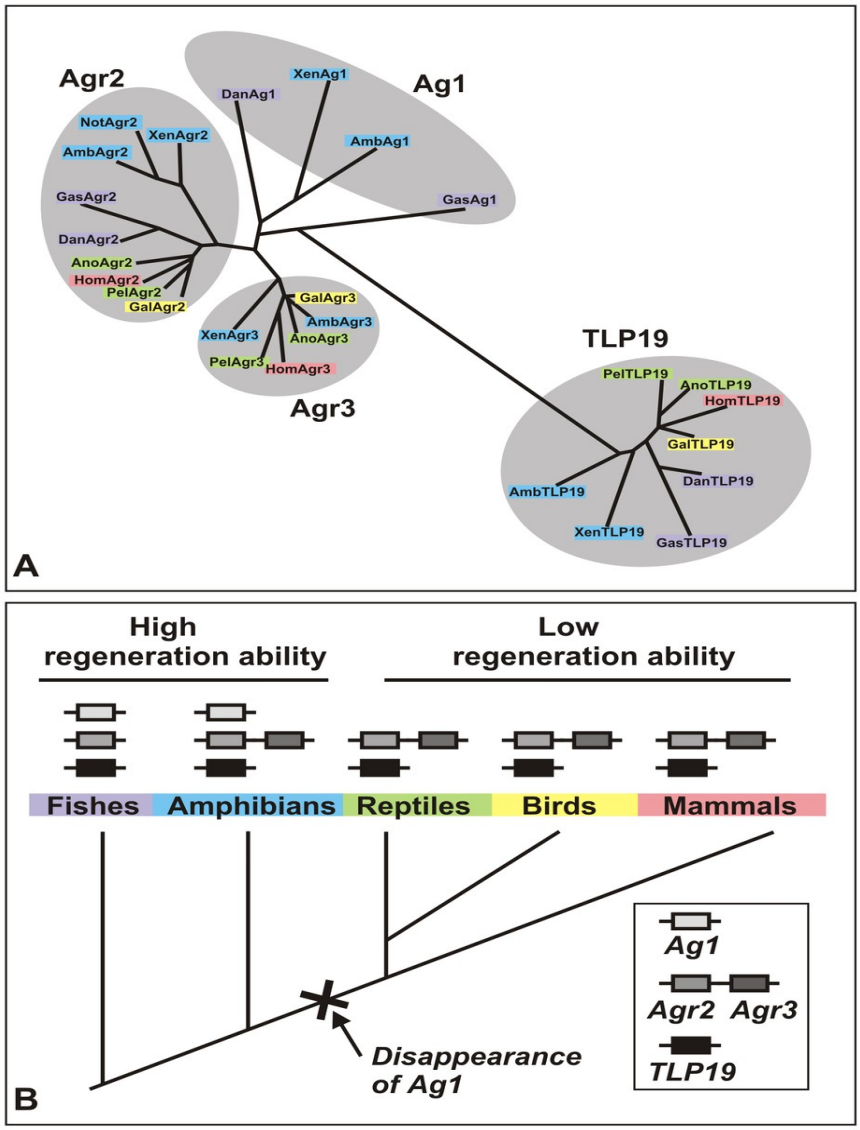


Figure 1.6 Phylogeny of *Agr* genes and evolution of different subfamilies of *Agrs* in vertebrates. (From [57]). (A) Analysis of local synteny of *Agrs*-containing loci in 17 amniotic and 6 anamniotic species with assembled genomes (from www.ensembl.org). The following species were analysed in this way: fishes (Danio rerio, Gasterosteus aculeatus, Oryzias latipes, Takifugu rubripes, Tetraodon nigroviridis); amphibians (Xenopus tropicalis); reptiles (Anolis carolinensis, Pelodiscus sinensis, Chrysemys picta bellii); birds (Gallus gallus, Meleagris gallopavo, Melopsittacus undulatus, Taeniopygia guttata); and mammals (Bos taurus, Gorilla gorilla, Homo sapiens, Ictidomys tridecemlineatus, Macropus eugenii, Monodelphis domestica, Mus musculus, Ornithorhynchus anatinus, Pan troglodytes, Sus scrofa). B) Evolution of *Agrs* subfamilies showed *Agr2* and *Agr3*, are present only in amphibians and higher vertebrates. *Ag1* are apparently absent in reptiles, birds and mammal but are present in fishes and amphibians.

1.3.3. Human AGR2

1.3.3.1. Gene architecture and expression of AGR2

The human *agr2* (*hAgr2*) gene spans a region of 50 kb in genomic DNA and is located on chromosome 7p21.3 [69]. Two mRNA isoforms were detected and contained eight exons and seven introns, and both isoforms contain the same open reading frame [70]. The sequence of both cDNA clones revealed that they have the same open-reading frame and exhibits an additional 5' exon. Zheng *et al.* investigated the promoter activity of hAgr2 gene using a luciferase reporter gene construct driven by the *agr2* promoter assay [71]. The results indicated that the forkhead box transcription factors FOXA1 and FOXA2 which maintains goblet cell function increased luciferase activity thus may regulate the transcription of *agr2* mRNA [71]. Additionally, the *agr2* promoter region also exhibits a binding site for hepatic nuclear factor 1, related to the same family as FOXA1 and FOXA2 [71]. Further work by Krig and co-workers using chromatin immunoprecipitation (ChIP) and database mining approach identified binding sites for the oncogene ZNF217, and relevant NF-κB and SOX9 within the agr2 promoter region [72].

AGR2 promoter regulations also have been implicated in hormonal response. AGR2 was significantly upregulated in response to estradiol [73] using both cell and animal models and showed a time and dose dependent manner induction by androgens [70]. Moreover, AGR2 can be regulated by pathophysiological stresses. When breast cancer cell line is subjected to serum starvation and oxygen deprived, *agr2* mRNA transcription increased approximately fivefold [74]. Several known inhibitors for ERK1/2, JNK, p38 and PI3K regulatory pathway was tested in this model, but only ERK1/2 inhibition resulted in reduced AGR2 mRNA expression using RT-PCR, suggesting that *AGR2* contributes to the survival of cells undergoing physiological stress in the microenvironment of the tumor. The *AGR2* induction response could be blocked effectively by a synthetic inhibitor of ERK1/2, suggesting that the activation signal for *AGR2* induction during stress is mediated through an ERK1/2-dependent pathway. AGR2 expression is controlled by the UPR upon ER stress and most likely depends on both IRE1 and ATF6 signaling [75]. IRE1α and ATF6α knockdowns resulted in decreased basal *Agr2* mRNA expression using RT-PCR and also prevented

Agr2 induction upon ER stressor, Tunicamycin treatment (an inhibitor of N-linked glycosylation). In contrast, PERK knockdown had no effect on *Agr2* expression. The study also showed that *Agr2* expression was partly under the control of the IRE1α-Xbp1s pathway where *Xbp1* mRNA splicing was attenuated in IRE1α dominant negative expressing cells. Other ER stress inducers like a reducing agent dithiothreitol (DTT) and thapsigargin (calcium pump inhibitor), were also found to upregulate agr2 transcription [70, 75] which supports the idea of the gene as protein disulphide isomerase (PDI) in maintaining ER homoeostasis. Further, AGR2 has been shown to be associated with tumour growth factor-β (TGF-β) responsive in pancreatic cells, whose expression can be downregulated by SMAD4 induction [76].

1.3.3.2. AGR2 protein structure

The AGR2 protein consists of 175 amino acid residues (Figure 1.7A), with a predicted molecular mass of 19,979.2 Da and a predicted pI of 9.03 [3]. AGR2 gene consists of 7 transcripts in total as shown in Ensemble[4]. However, only 4 of them has the open reading frame and protein coding: AGR2 001, AGR2 005, AGR2 006, and AGR2 007, where AGR2 001 is commonly used as the reference sequence. A sequence alignment of AGR2 reveals 54% identity and 71% similarity to XAG-2, and 91% identity and 96% similarity to murine AGR2 [77]. The availability of bioinformatics software to predict structural motif highlights several important characteristics of the expressed protein. Based on the secondary structure prediction (PSIPRED v 3.3)[5] the full-length AGR2 was shown to possess α- β-α- β- α- α- β- β- β- α arrangement (Figure 1.7B). The N-terminal 45 amino acid residues exhibit high levels of predicted disorder using DISOPRED3 tools (Figure 1.7C) [78]. Regions of intrinsic disorder were described to particularly enriched in proteins involved in protein-protein interactions, cell signalling and transcription [79]. A more recent approach using the powerful Nuclear Magnetic Resonance (NMR) to study the crystal structure of mature AGR2 (without the 20-amino-acid signal peptide) revealed that N-terminal amino acid region 21– 40 is unfolded which was consistent with the predicted disorder region data [61]. The N-terminal unstructured region also showed an enhanced rate of adhesion of rat

[3] ExPASy Compute pI/Mw. (http://web.expasy.org/compute_pi/)
[4] http://www.ensembl.org
[5] PSIPRED v 3.3 http://bioinf.cs.ucl.ac.uk/psipred/

mammary tumour cells which supports the impression that intrinsic disorder region plays a role in protein's function. Structural analysis of $AGR2_{41–175}$ (PDB code: 2LNS) showed that it has α- β- α- α- β- β- β- α arrangement which was consistent with its secondary structure prediction (Figure 1.8A). Each AGR2 monomer subunit consists of a four-stranded β -sheet core surrounded by four α - helices: α 1 (residues 58– 67), β 1 (residues 72– 77),α 2 (residues 82– 93), α 3 (residues 95– 103), β 2 (residues 105–108), β 3 (residues 127– 132), β 4 (residues 135–138) and α4 (residues 157–168) [61]. The NMR structure analyses also showed that AGR241–175 exists as homodimer which occurs via antiparallel rearrangement of the α1 helix (Figure 1.8B). The 60-EALYK-64 within the α1 helix is shown to be the dimeric interface where the intersubunit salt bridge between the carboxylate group of E60 and the ammonium group of K64 forms the major interaction.

There are three major structural determinants of AGR2. First, AGR2 harbours hydrophobic signal peptide sequences in its N-terminus (amino acids 1-20) directing AGR2 into the endoplasmic reticulum. The putative cleavage sites for signalling peptides predicted by Signal-P program[6] is between amino acids Ala20 and Lys21. The N-terminal leader sequence target nascent polypeptides directly from the ribosome to the ER membrane via a signal recognition particle (SRP) and later cleaved by signal peptidase in the ER lumen, yielding a mature, processed form of a protein [16]. However, there is very little information on regulation at the level of cleavage of the signal peptides by enzymes such as signal peptide peptidase [80]. Cell-based analysis using immunofluorescence of localisation of AGR2 with and without signal peptide suggested differential subcellular localisation where the full-length protein is localising predominantly to the ER and the shorter, mature AGR2 lacking the N-terminal leader sequence localising primarily to the nucleus [81]. As described by Vitale and Denecke, co-translational cleavage by signal peptidase of protein harbouring signal peptide is essential for protein folding and maturation [16]. Thus protein with an absence of the signal peptide would not be directed to the ER lumen, and instead, would demonstrate subcellular mislocalisation. Further, it has been indicated that signal peptide peptidase can cleave intramembrane signal sequences

[6] http://www.cbs.dtu.dk/services/SignalP/

which can lead to perturbation of endoplasmic reticulum-associated protein degradation (ERAD) to control the unfolded protein response (UPR) modulators [82]. Studying whether AGR2 involves in proteolysis by signal peptide peptidases and impact upon endoplasmic reticulum quality control, unfolded protein responses, and disease mechanisms would gain more insights in AGR2 signalling.

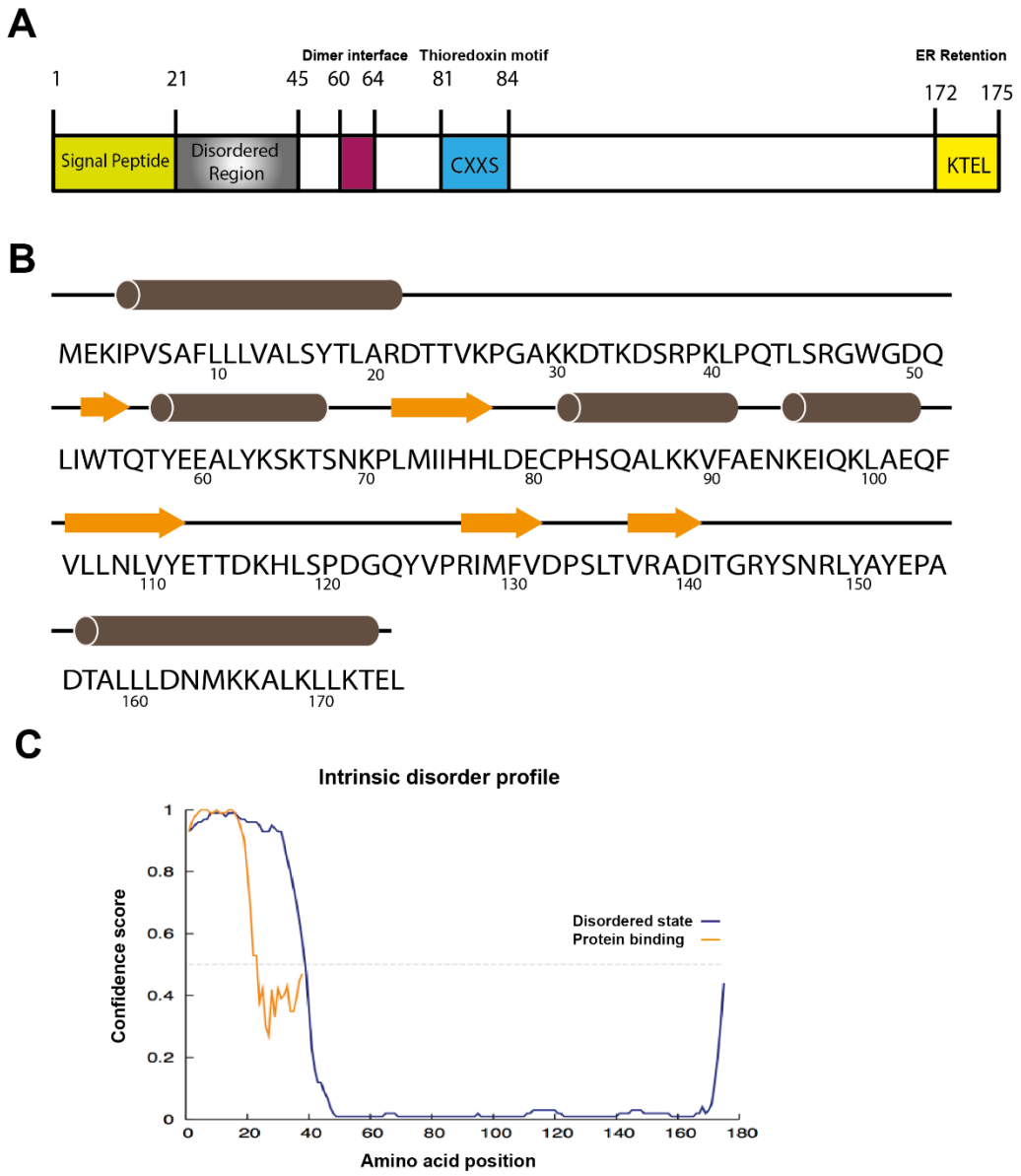


Figure 1.7 Structural elements of AGR2. (A) Schematic illustrating structure of the full-length AGR2 protein. The panel highlights the functional motifs human AGR2; signal peptide, an N-terminal intrinsically disordered region; a dimerization motif; a thioredoxin fold, and the ER retention site. (B) Schematic illustrating predicted secondary structure of AGR2 (PSIPRED v 3.3) (C) The N-terminal 45 amino acid residues indicate a high level of predicted protein disorder using DISOPRED3 [78].

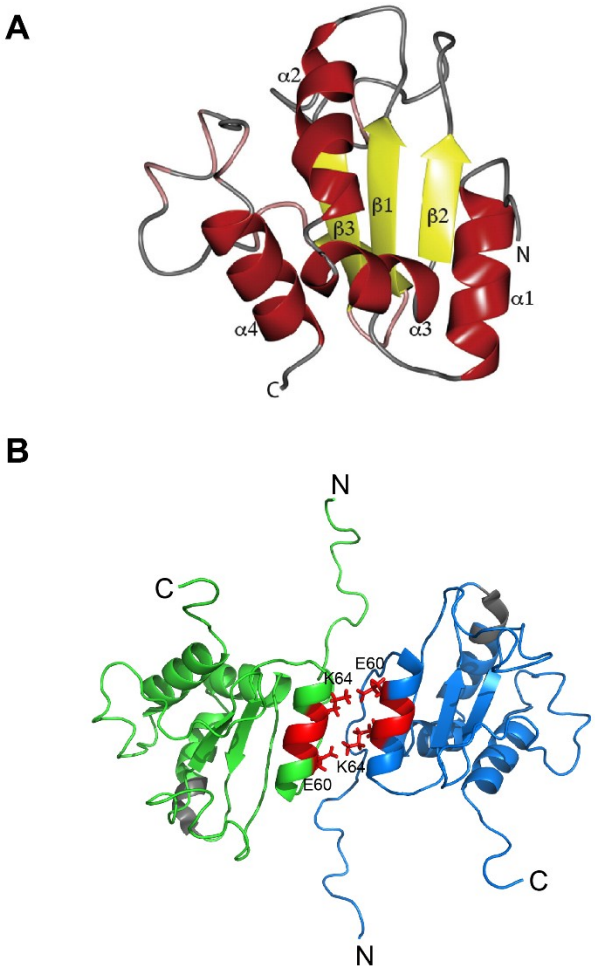


Figure 1.8 Crystal structure of AGR2$_{41-175}$ (PDB code: 2LNS). A) Ribbon representation of the subunit structure of AGR2$_{41–175}$ using NMR showing secondary structures (β-sheet in yellow and α-helix in red) (reproduced from [61]). B) Ribbon representation of homodimer model of AGR2$_{41–175}$ showing the antiparallel arrangement of the α1 helix. Dimeric interface 60-EALYK-64 is highlighted in red with the residues of E60 and K64 showed as sticks. The intersubunit salt bridge between the carboxylate group of E60 and the ammonium group of K64 forms the major interaction between the subunits. Ribbon representation were generated using PyMOL v1.8.0

The second key motif of AGR2 primary structure is the extreme C-terminus ER retention motif containing tetrapeptide sequence of lysine (K) – threonine (T) – glutamic acid (E) – leucine (L) abbreviated as KTEL [83]. This motif is a non-optimized form of the classic ER retention motif, whose classical form is H/KDEL [84]. More recent studies have suggested the role of extra amino acids at position 5 and 6 from the C-terminus in determining ER residency [85]. The ER retrieval motif KDEL, complemented with N-terminal signal peptide, also targets proteins to ER, where it mediates retrograde transport of proteins in endoplasmic reticulum during the protein secretion process into the extracellular space [85]. Early studies showed that proteins possessing H/KDEL or similar motif interact with Golgi-resident KDEL receptors in intermediary compartments of cis-Golgi complex [86]. Having H/KDEL been well described as an ER retrieval motif, a few examples of human proteins exist where proteins exhibiting this motif demonstrate a non-ER localisation such as the cytosolic and nuclear localisation of the KDEL-containing protein calreticulin [86]. The underlying mechanism of how these proteins can evade ER retention remain unknown, whether it is through an error rate of the KDEL receptor molecules, active secretion or whether KDEL-expressing proteins are released by necrotic cells. Variations in the sequence have been described [66, 83], including KTEL which is possessed by AGR2, which has reduced binding affinity for the KDEL receptor [85, 86]. Hence, this might explain why KTEL motif of AGR2 provides some flexibility in the subcellular localisation where it can 'leak' outside the ER, which may be essential for dynamic protein function in response to physiological conditions [81, 83]. Modulating the motif in AGR2 has shown effects in cell-based systems where deletion of KTEL result in secretion of AGR2, and while KTEL mutagenesis to the optimised canonical KDEL increases the magnitude of endoplasmic reticulum localisation [81]. Gupta *et al*. also showed that AGR2 induce the expression of amphiregulin (AREG) only when the wild-type carboxyl-terminal KTEL only with the highly conserved KTEL but not with KDEL [83]. Hence, variants of ER retention sequence may serve a specific functional role, and in the case of AGR2, this role is served specifically by KTEL. In addition, deletion of the motif eliminates its ability to stimulate cell growth in clonogenic assays and to attenuate the p53 transcriptional response to DNA damage [87] which suggest an AGR2 function in cancer cell survival that is just not limited to

localisation in the ER. Also, a recent study showed surprising results. In newt AG (nAG), when myc-tag is introduced at its C-terminal, the nAG protein after transfection into cells was highly secreted into media which was observed by chance by Grassme *et al.* [68]. When the proteins were further extended for 25 residues after the myc-tag, the levels of secretion were even higher. This maybe due to the KTEL is hidden with the addition of the extra myc-tag which may have a lower binding affinity to KDEL receptor, and therefore the protein is not retained inside the cell but rather secreted into the media. The newt protein nAG was identified as a binding partner of the salamander-specific protein Prod1, a member of the three-finger protein (TFP) superfamily that is GPI-anchored at the cell surface. secreted nAG acts on blastemal cells via the GPI-linked Prod1 receptor [4]. nAG acted to promote S phase entry of cultured newt blastemal cells, possibly by binding to Prod1 on the cell surface. This study showed that the wt nAG and its extended variant (myc+25) had comparable activity in promoting S phase entry of cultured blastemal cells.

Another key structural motif of AGR2 which is shared with its orthologue AGR3 is the CXXS motif or thioredoxin fold which taxonomically classify the protein as a member of protein disulphide isomerase (PDI). Thioredoxin fold is classically described as having the CXXC motif with dual cysteines mediate oxidoreductase which is important in shuffling disulphide bonds during protein maturation [66]. Having the sole cysteine residues in CXXS motif of AGR2 possesses lower oxidoreductase activity but may, however, contribute to isomerization of already existing disulphide bridges and possibly perform specialised functions in the ER [88, 89]. Other oxidoreductases also harbour CXXS motifs including ERP44 [90, 91]. The founding gene of the AGR family exists in invertebrates and is named ERP18 (also named AGR1) [92, 93]. This protein contains the classic di-cysteines thioredoxin fold with the CxxC motif and demonstrates through the covalent trapping of disulphide bonds in ER [93]. Though ERp18 and AGR2 are highly similar along with having the thioredoxin fold, ERP18 did not emerge as oncogenic signalling molecule using OMICS discovery platforms like AGR2 [94]. Interestingly, studies using transgenic mice showed different phenotypes between AGR2 knockout mice compared to its ortholog AGR3; where animals in the absence of AGR3 showed lower ciliary beat

frequency in the airway [95]. This indicates the non-reductant function of AGR2 and AGR3 function in development and gives a hint that the single cysteine in AGR2 is vital in its substrate binding succession. Surprisingly, Bergstrom *et al.* have recently demonstrated that in addition to the KTEL ER-retention signal, the presence of a single cysteine within the AGR2 thioredoxin-like domain is required for the control of protein secretion [96].

Apart from these three classic biochemical determinants that drive AGR2 functions, recent studies have brought another compelling insight about AGR2 protein architecture. The NMR studies showed $AGR2_{41\text{-}175}$ can exist in homodimer structure through a novel amino acid motif at position E60-K64 in an antiparallel arrangement of the α1 helix [61]. The salt bridge between the carboxylate group of Glu60 and the ammonium group of Lys64 forms the major interaction between the subunits in an anti-parallel fashion. Mutagenesis at this dimerization motif abolished the dimeric structure. Sequence alignment showed that E60 is conserved in all 21 species from zebrafish to humans, and K64 is conserved in amphibians, marsupials and mammals but as R in mouse and, in general, tryptophan in fish [61]. It is important to note that the dimerization interface of AGR2 is held far from the catalytic CPHS motif (opposing faces), where enhanced disulphide exchange in client proteins might occur. Another study showed that the N-terminal disordered region (amino acids 20-40) is another determinant in AGR2 dimeric structure where deletion in that region significantly stabilises the dimer [97]. Remarkably, AGR2 can form an alternatively shaped dimeric structure in which oxidation-dependent homo-dimerisation occurs through Cysteine-81 mediated disulphide bond formation which would re-orientate the dimer into a different conformation [98]. Hupp group also contributed to the molecular arrangement of AGR2 in a recent study. The study showed that mutation of Cysteine-81 to alanine prevents peroxide catalysed dimerization of AGR2 in vitro, suggesting a reactive cysteine is central to covalent dimer formation. Using an ESI mass spectrometry Clarke *et al.* demonstrated that low levels of a chemical oxidant promote an intermolecular disulphide bond through the formation of a labile sulfenic acid intermediate [99]. However, higher levels of oxidant promote sulfinic or sulfonic acid formation thus preventing covalent dimerization of AGR2. Thus single cysteine (C81)

of the AGR2 act as an oxidant-responsive moiety that regulates its tendency for oxidation and its monomeric-dimeric state. Physiologically, this suggests that the monomer to dimer equilibrium is important in AGR2 signalling and dictates the protein geometry and synergy between the dimer interface and the catalytic thioredoxin motif given the fact that AGR2 has only sole Cysteine-81 for thiol-disulphide interchange to assemble correctly folded disulphide bridges in client cargo proteins**.**

1.3.4. Functions of AGR2

Functions of AGR2 are not well defined since AGR2 is not only found in ER but also in other subcellular localisation such as nucleus [81, 83], cell surface and extracellular space [100]. Human AGR2 is strongly expressed in tissues of the lung, stomach, colon, prostate and small intestine, organs which contain mucus-secreting cells, or function as endocrine organs (Table 1.2) [101]. Likewise, human AGRs are predominantly expressed in endoderm-derived organs [102]. In normal tissues, AGR proteins could contribute to the regulation of the total protein load in the cell. For instance, in normal mammary gland AGR2 highly expressed during late pregnancy and lactation [103], a period during which high volumes of milk protein are produced in the secretory alveolar buds. During this period, AGR2 promotes epithelial cell proliferation and milk protein production from the assistance of additional protein-folding and secretory factors to cope with the increased protein load. Moroever, milk protein expression at the mRNA level is downregulated in a mammary-specific Agr2 knockout mouse model and upregulated in a mammary-specific Agr2 overexpression mouse model [103], suggesting that AGR2 may play a role as a gatekeeper at both the protein and mRNA level to regulate the total protein load a cell can withstand without entering proteotoxic-induced cell death pathways. Another line of evidence is thatAGR2 is required for the production of airway epithelial MUC5AC and MUC5B in respiratory system tissues after allergen challenge [104].AGR2 may also be involved in epithelial barrier function since the AGR2 promoter regulators are grouped into a family which are typical for epithelial goblet cells [71]. Here, possible roles of AGR2 function based on its different localisation and its interacting proteins are described.

Organ system	Normal tissues IHC	Cell type	Protein expression
Liver and pancreas	Prostate	Glandular cells	Strong
	Gallbladder	Glandular cells	Strong
Digestive tract	Stomach	Glandular cells	Strong
	Duodenum	Glandular cells	Strong
	Small intestine	Glandular cells	Strong
	Appendix	Glandular cells	Strong
	Colon	Glandular cells	Strong
	Rectum	Glandular cells	Strong
Urinary tract	Urinary bladder	Urothelial cells	Strong
	Kidney	Cells in tubules	Moderate
Male reproductive system	Epididymis	Glandular cells	Strong
	Prostate	Glandular cells	Strong
	Seminal vesicle	Glandular cells	Strong
Breast and female reproductive system	Breast	Glandular cells	Weak
	Cervix, uterine	Glandular cells	Strong
	Fallopian tube	Glandular cells	Strong
	Endometrium	Glandular cells	Strong
	Placenta	Trophoblastic cells	Moderate
Skin and soft tissues	Skin	Melanocytes	Weak
Blood and Immune System	Tonsil	Squamous epithelial cells	Moderate
	Cerebral cortex	Neuronal cells	Moderate
Respiratory system	Lung	Alveolar cells	Moderate
	Nasopharynx	Respiratory epithelial cells	Strong
	Pancreas	Exocrine glandular cells	Moderate
	Tonsil	Squamous epithelial cells	Moderate

Table 1.2 AGR2 expression in normal tissues (derived from Human Protein Atlas[7] as described in [101]). IHC= Immunohistochemistry.

[7] http://www.proteinatlas.org

1.3.4.1. Interactomics of AGR2 defining its function

Identifying the interacting proteins (termed interactomics) is one of many ways to understand how the protein works. Several studies have used a variety of techniques to uncover the protein interaction network of AGR2 (Table 1.3). Yeast-two hybrid (Y2H) and co-immunoprecipitation, two commonly used assays to study protein-protein interaction, have largely driven the comprehensive discovery of AGR2 binding proteins. Studies using Y2H that human AGR2 can interact with the metastasis-associated GPI-anchored C4.4a protein and extracellular alpha-dystroglycan (DAG-1) [58] and previously described membrane receptor Prod1 [54]. These data showed that AGR2 plays a pivotal role in the secretion of the receptor protein that is linked to adhesion and cell growth. Further Y2H analyses shown that AGR2 interactome is not clouded to only membrane receptor but to nuclear binding proteins such as Chromodomain Helicase DNA Binding Protein 6 (CHD6) and Reptin [94, 105]. Reptin is the AAA+ ATPase protein and the only AGR2 client protein that has been well studied [105]. The AGR2: Reptin interaction was mapped to a peptide interface, comprising the proposed substrate-binding loop on AGR2, within amino acids F104 to Y111. Amino acid mutagenesis at that positions attenuates Reptin binding to the AGR2. Later it was shown that Reptin participated in the cytosolic regulatory protein-folding complexes through the chaperone-dependent R2TP activity [106] and in cilia complex assembly at the plasma membrane [107] which could drag the involvement of AGR2. Studies thus far assumed that AGR2 acts as a monomer, but recent NMR studies recently showed that AGR2 is predominantly in dimeric form with a dissociation constant 8.83 μM where high concentrations will be maintained in its resident location, ER. There is very little information about the dimeric function of AGR2 towards binding to its client protein. Gray *et al.* work showed that AGR2 in this condition has gain-of-function towards binding to its client protein. The study revealed that a stable dimer Δ45-AGR2 was more active in binding to Reptin than either wt-AGR2 or the monomeric AGR2 (E60A) mutant [97]. Another study by Ryu *et al.* showed that AGR2 dimerization is required for the interaction with BiP/GRP78 in mediating the unfolded protein response (UPR) signalling pathway in ER [98]. In addition, it has been shown that PDI oligomerization was associated with chaperonin activity independent of isomerase activity [108] which could foreshadow dimeric

AGR2 function. Therefore, as also discussed in AGR2 protein structure1.3.3.2, studying the relationship between the equilibrium of monomer to a dimer of AGR2 and its interactome will open the door for more research niche in the future

Although CoIP is considered to be the gold standard assay for protein-protein interactions and has been the driving force in determining AGR2 client proteins, there are some limitations for this assay. First, the interaction partners may not bind directly, since they can be part of larger complexes thus may interact via one or more bridging molecules [109]. Second, co-precipitation is insensitive compared to other methods such as protein affinity chromatography, because the concentration of the antigen is lower than it is in protein affinity chromatography. There are also several setbacks for Y2H system. For example, Y2H is limited to proteins that can be localized to the nucleus, which may prevent its use with certain extracellular proteins [110]. The proteins must also exist and stably expressed in yeast cells and to retain activity as fusion proteins for detection. Furthermore, the environment of the yeast nucleus inhibits interactions of some proteins. The use of protein fusions also means that the site of interaction may be blocked by one of the transcription factor domains. In addition, posttranslational modification that does not occur in yeast cells therefore interaction that is dependent on this cannot be detected. While both CoIP and Y2H system as well as other protein-protein interaction methods can identify real protein-protein interactions, they cannot tell where does the interaction occur in the cell. Several new techniques of identifying client proteins and ways to overcome these drawbacks will be discussed later in the thesis which can further expand AGR2 interactome.

Gene name	Function	Method	Reference
DAG1	Links the cytoskeleton and the extracellular matrix	Y2H	Fletcher *et al.* (2003) [58]
REPTIN	AAA+ ATPase—DNA repair and transcription	Y2H	Maslon *et al.* (2010) [105]
KDELR	KDEL receptors	CoIP	Raykhel *et al.* (2007) [86]
C4.4	Metastasis-associated GPI-anchored protein	Y2H	Fletcher *et al.* (2003) [58]
MUC2	Mucin 2	CoIP	Park *et al.* (2009) [60]
PROD1	Axolotl homolog for human CD59	Y2H	Kumar *et al.* (2007) [54]
EGFR	Epidermal Growth Factor Receptor	CoIP	Dong *et al.* (2015) [111]
ARGHAP29	GTPase-activating protein	Y2H	Unpublished data [94]
CKAP2	Cytoskeletal-linked protein involved in mitosis	Y2H	Unpublished data [94]
CHD6	Chromatin-remodelling factor	Y2H	Unpublished data [94]
GPSM2	Regulates G-protein activation	Y2H	Unpublished data [94]
HECTD1	E3 ubiquitin ligase	Y2H	Unpublished data [94]
LYPD3	Regulates cell migration	Y2H	Unpublished data [94]
HIVEP1	DNA binding protein	Y2H	Unpublished data [94]
NRIP1	Binds hormone-dependent receptors	Y2H	Unpublished data [94]
NRXN3	Controls adhesion and receptor signalling	Y2H	Unpublished data [94]

Table 1.3 Published AGR2-interacting proteins identified from classical protein-protein interaction methods. (Modified from [94]) Y2H= Yeast two-hybrid, CoIP=Co-immunoprecipitation.

1.3.4.2. AGR2 as an endoplasmic reticulum protein disulphide isomerase

AGR2 has been classified as a member of PDI family of molecular chaperone due to the presence of thioredoxin-like fold based on phylogenetics analyses [66]. Coupled with the degenerative ER retrieval motif and hydrophobic N-terminal signal peptide, it is obvious to say that ER is AGR2 native resident where it is most functional. Though AGR2 functions have been both speculated intra and extracellularly, its protein disulphide isomerase function might be pertinent as a protein folding catalyst. The ER contains a network of chaperones which processes proteins targeted for membrane insertion or extracellular secretion by performing folding, disulphide bond introduction, and N-glycosylation through distinct complementary pathways. The ER consists of chaperones that belong to several classical chaperone families, such as heat shock protein (Hsp) 40, Hsp70 and Hsp90 [15]. The ER also houses unique chaperones and folding enzymes such as calnexin and calreticulin and the family of PDIs and Peptidyl-prolyl isomerases (PPIases) [15]. These chaperones and folding enzymes are part of the ER quality control (QC) system that ensures protein are properly folded before entering the subsequent secretory pathway.

The main function of the PDI family members is to catalyse the formation or rearrangement of the disulphide bonds between two cysteine residues of nascent proteins in the lumen of the ER for their proper folding and maturation prior to the release for cellular transport [112]. The presence of the canonical thioredoxin fold, CXXC, contains two free thiol groups at each of the cysteine residues allows these oxidoreductase activities to take place. The thioredoxin fold is thought to be the minimum requirement for PDI family membership. Most of PDI family members have between one to three dithiol redox motif (CXXC) [113]. To date, 21 members varying in size, structure, tissue distribution and enzymatic activity were identified [114]. AGR2 is a class of PDI with one divergent CXXS (CPHS) active site sequence which means it lacks the fully oxidoreductase CXXC motif activity. It is notable that the cysteine to serine substitution is common for some PDI family members, however, the activity of CPHS active site is unknown [115]. Some PDIs also has chaperone activity to fold denatured proteins which are independent of its thiol activity. This activity is

thought to occur via the exposed hydrophobic residues, intrinsically unfolded loops, or conceivably exposed aromatic residues which suggest substrate specific client proteins are required [113].

AGR2 has been described as a dominant component of ER homoeostasis [75]. AGR2 was identified in proteomics analysis of ER-bound ribosome that bound to newly synthesised cargo proteins. Further, it was shown that AGR2 expression could be controlled by the unfolded protein response (UPR) and that silencing of AGR2 disrupts the components of the ER-associated degradation machinery (ERAD), resulting in a reduction in cells ability to cope with acute ER stress [75]. This showed that AGR2 can act as a switch, whether it can act as a catalyst in protein folding of nascent proteins or involve in control ERAD which are both important in the stringent requirement of ERQC exit [116] (Figure 1.9).

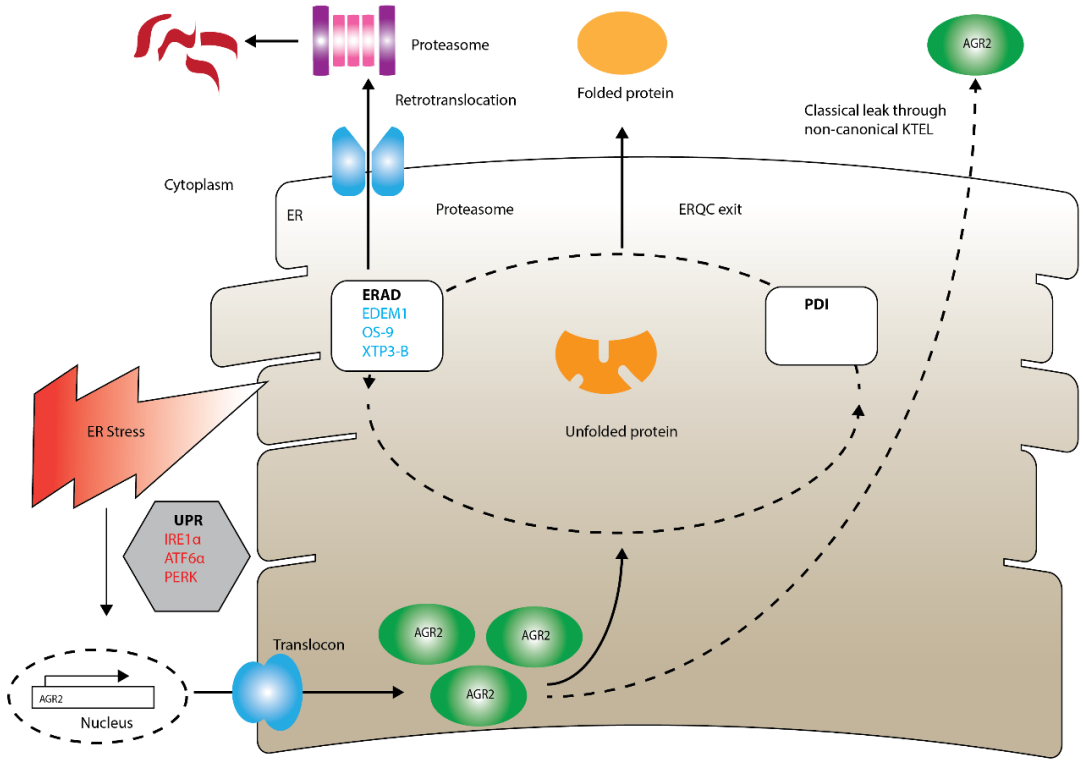


Figure 1.9 Functions of AGR2 in the ER. (From [116]) AGR2 can respond to classic unfolded protein response pathways controlled by PERK, ATF6, and IRE1 or can act through ERAD pathways. However, co-factors through which AGR2 can facilitate receptor maturation and secretion are yet to be investigated. The AGR2 expression is elevated in response to ER stress in disease landscape like cancer that demands mutant protein production and protein synthesis. AGR2 may also act outside the ER (classical leak due to possession of non-canonical ER retention motif) through an unknown mechanism.

1.3.5. AGR2 associated diseases

1.3.5.1. Asthma

AGR2 gene has been discovered as a biomarker for asthma in murine models of acute and chronic asthma [117]. Gene expression microarray in the lung of these models and confirmed with RT-PCR identified AGR2 as one of the overexpressed genes. Immunohistochemistry showed that AGR2 primarily localised to the goblet cells of the lung of these models [117]. Further study in human lung epithelial cells exhibited that knocking down interleukin-13 (IL-13), an "inducer" that causes the proliferation of goblet cells and over-production of mucus in the air- way, had a downstream effect on AGR2 and MUC5AC levels through a signal transducer and activator of transcription 6 (STAT6)-dependent pathway [118].

These results were succeeded by the study to determine the relationship of AGR2 with other airway mucins, namely MUC5AC and MUC5B [104]. AGR2 was shown to colocalise in the same cells as MUC5AC and MUC5B and was able to co-immunoprecipitate with immature form of MUC5AC. AGR2 was found to express in the ER of MUC5AC- and MUC5B-producing mucus cells in the airway epithelium and submucosal glands. Mice deficient in AGR2 (*agr2* -/-) have impaired production of MUC5AC and MUC5B as well as MUC2. Interestingly, AGR2 mice had compromised export of mucins which found accumulated in the ER that in turns activate the unfolded protein response. These findings implicated AGR2 role in the elevating production of respiratory mucins MUC5AC and MUC5B responsible for mucus production. Consequently, overproduction of mucus results in airway closure that is one of the prominent features of asthma [119, 120]. Further studies also demonstrated that AGR2 expression was higher in asthmatic mice and showed a positive linear relationship with the expression of IL-13 and MUC5AC [118]. Dexamethasone, a glucocorticoid drug induced silencing of IL-13 attenuated AGR2, MUC5B, and IL13 levels in epithelia of the airway suggesting a direct role of AGR2 in disease penetrance [121].

1.3.5.2. Inflammatory bowel disease

Inflammatory bowel disease (IBD) is a group of autoimmune disorders of the gastrointestinal tract. This disease is categorised into two principal: Crohn's disease (CD) and ulcerative colitis (UC). The chromosomal location of AGR2 and its homologue AGR3, 7p21.3 [69], is mapped to a key susceptibility region of inflammatory bowel disease in a genome-wide linkage analysis [122, 123] which lead to the investigation of AGR2 as a candidate gene for inflammatory bowel disease [71]. It was demonstrated that single nucleotide polymorphism (SNP) were associated with the 5' region of the AGR2 gene to UC phenotype in two independent populations (The UK and German cohort) [71]. The expression level of AGR2 in UC patients was significantly lower than in healthy controls. In addition, carriers of the risk alleles also showed lower expression of AGR2. The study also found that AGR2 promoter is activated by forkhead box transcription factors FOXA1 and FOXA2 responsible for goblet cell function maintenance. These findings are consistent in the study of transgenic mice described above, where the mice that are null for AGR2 found to be susceptible to ulcerative colitis [60, 63]. Further, the intestines of patients suffering from inflammatory bowel disease often exhibit significant symptoms of ER stress [124] [125], suggesting that dysregulation of the ER stress response associated with low levels of AGR2 may pose a secondary pathway for increased susceptibility to IBD. Altogether these results support the hypothesis that AGR2 involve in maintenance epithelial barrier integrity [124, 125].

It seems that diseases associated with AGR2 is completely overlapped with those associated with UPR diseases as UPR has been implicated in diseases similar to AGR2 discussed above . However the scope UPR-associated is broader which include metabolic disease, and neurodegenerative disease. One example that links AGR2 and UPR is that in the AGR2 knockout mice studies which showed that there is a decreased expression of MUC2 in goblet cells and abnormal localization of Paneth cells in the small intestine, which overlaps with UPR activation and the development of spontaneous ileocolitis [63].

The role of AGR2 in the secretory pathway, and relief or ER stress, reviewed above, support the finding that in the normal functioning of the cell controlled AGR2 expression is a requirement for the appropriate maturation, localisation and secretion of extracellular proteins. AGR2 alteration through mutation, overexpression and underexpression has indicated significant pathological effects.

1.3.6. AGR2 in cancer

1.3.6.1. AGR2 pro-oncogenic features

Human AGR2 gene was first described in the oestrogen receptor (ER)-positive MCF-7 cell line [126]. Regulation of AGR2 protein expression by oestrogen was later confirmed by both *in-vivo* [73] and *in-vitro* experiments [127]. AGR2 also was found in other types of cancer pertaining to hormone-dependent such as [58, 74, 101, 103, 126-132] prostate [70, 101, 133, 134] and ovarian [135-138], as well as in non-hormone dependent cancers such as tumours of the oesophagus [87, 139], gastrointestinal tract [140] and lungs [141, 142]. AGR2 exhibits the basic features of a pro-oncogenic protein and cell survival enhancing activity. Clonogenic assays were used to demonstrate that AGR2 overexpression can enhance cancer cell survival, rather than inhibit cell growth in the premalignant Barrett oesophagus and oesophageal cancer models [87]. Deletion of the last 10 amino acids of AGR2, harbouring the ER retention site of AGR2, prevented clonogenic growth stimulation and influencing p53 phosphorylation indicating a role in tumourigenesis of the ER localised protein and survival factor. Further, AGR2 silencing was shown to inhibit proliferation, invasion and survival *in-vitro* in pancreatic [143] and breast cancer cell lines [144]. Moreover, conditioned media from cells silenced in AGR2 have a reduced ability to stimulate proliferation of pancreatic cancer cells [143]. On the contrary, short interfering RNA- or short hairpin RNA-mediated AGR2 knockdown prevents colony and subcutaneous growth in oesophageal and pancreatic cancer models [62, 143].

1.3.6.1. AGR2 overexpression in cancers

Large-scale cancer genomics data sets[8] and as described in [145] showed that most AGR2 were amplified in cancer landscape (Figure 1.10). Only a few mutations were reported in the dataset which mainly exhibited missense mutations (one amino acid change), suggesting that mutants AGR2 had little effect in cancer development. In human, AGR2 had previously been first identified as an overexpressed gene in breast cancer cell lines [126] and breast cancer patients especially in association with oestrogen positivity [58]. In fact, the top literature findings using data from Pubmed[9] showed that AGR2 was most associated with breast cancer. The overexpression in breast cancer is clinically relevant as high levels of AGR2 are associated with poor prognosis in ER-positive breast cancer [127, 146]. The expression of AGR2 in poor-prognosis tumours may be due to multiple regulatory mechanisms whether an AGR2 expression can be induced in an ER-dependent manner early in tumorigenesis, leading to an inherent therapeutic resistance, or its expression can be induced in response to anti-hormone treatment, leading to an acquired resistance to therapies. Furthermore, the high AGR2 expression is sometimes found in ER-negative breast cancers [58], indicating a different mechanism of upregulation, such as physiological stress [74]. Breast tumours can be classified into the basal-like, Her2, luminal A, luminal B, and normal breast-like subtypes by their intrinsic gene expression profiles [147]. AGR2 is expressed significantly higher in the ER-positive luminal A and B subtypes compared with the ER-negative basal-like, Her2, and normal-like subtypes, supporting the previously reported overexpression of AGR2 in ER-positive breast cancer [146].

[8] http://www.cbioportal.org/
[9] http://www.cancerindex.org/geneweb/AGR2.htm

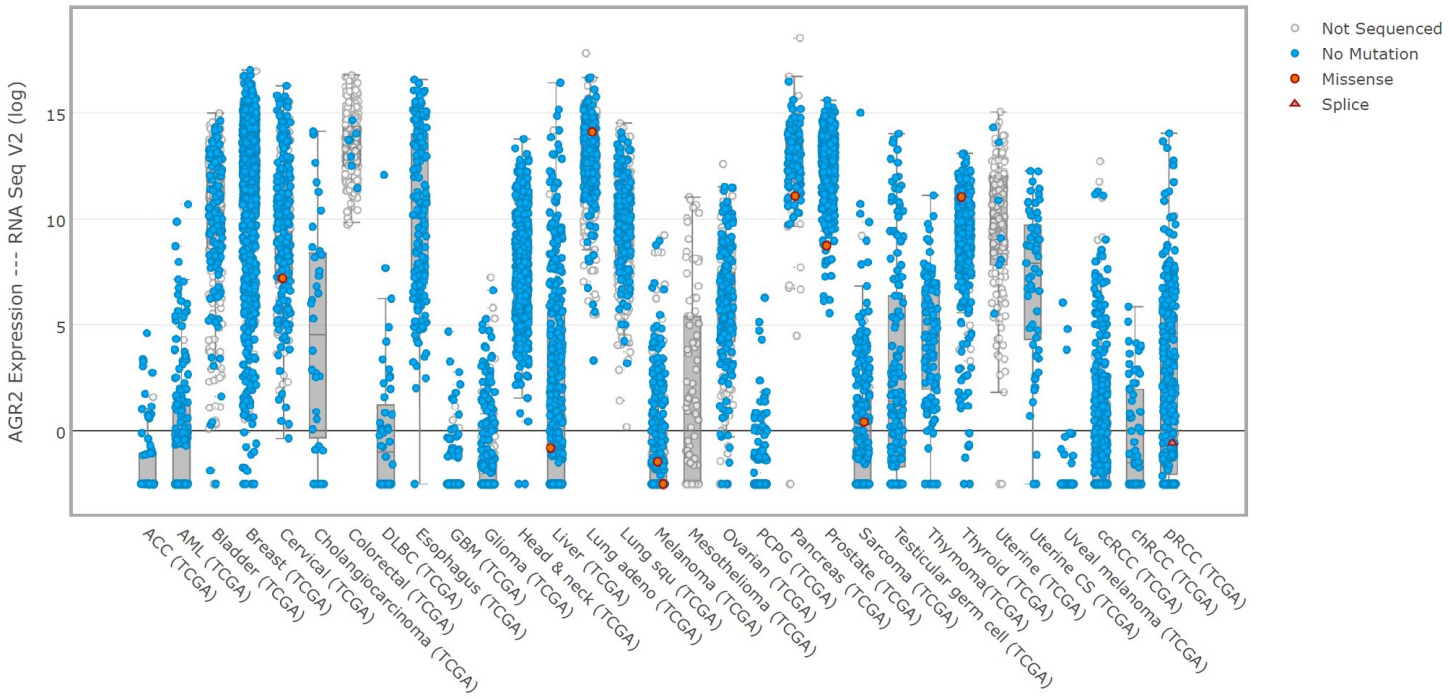


Figure 1.10 AGR2 expression from cBioportal large-scale cancer genomics data sets as described in [145]. This portal contains data from 147 cancer genomics studies that reduces molecular profiling data from cancer tissues and cell lines into user-friendly and interactive genetic, epigenetic, gene expression, and proteomic events. Data showed the top genomic studies in AGR2 cases**.** A large number of cancers showed overexpression of AGR2 with few cases of mutations, where mostly missense mutation.

AGR2 has been also implicated in a various type of cancers as summarised in (Table 1.4). In ovarian cancers, mucinous cancer biopsies showed overexpression of AGR2 expression by immunohistochemistry [137], as well as in endometrioid cancers[148], and high-grade serous tumours [149]. A high concentration of AGR2 in the extracellular plasma has been associated with both serous and non-serous ovarian cancer [135]. In prostate carcinomas, AGR2 has been reported in high concentration [150, 151], both intracellularly and extracellularly in the blood [151] and urine [152], and AGR2 is now has been selected as possible candidate for a biomarker for prognosis of prostate cancer. Increased level of AGR2 is associated with poor prognosis of prostate cancer patients [134].

AGR2 overexpression is also associated with the normal biliary tree in both fetal and adult normal liver. Strong AGR2 expression was seen in the tall epithelial cells covering the large bile ducts as well as gallbladder epithelial cells [153].

Elevated levels of AGR2 have also been linked to non-hormone related oncogenesis and cancer progression, and this includes oesophageal adenocarcinoma. Oesophageal adenocarcinoma is different from other types of cancer where the environmental stress of bile acid reflux plays a seemingly important role in disease progression [154]. In addition, p53 tumour suppressor gene mutation can occur very early in the disease sequence; during the conversion of squamous epithelium with metaplastic epithelium, also called ‘Barrett’s oesophagus’. Proteomics screen in the preneoplastic Barrett’s epithelium identified AGR2 as overexpressed proteins, which was validated as a strong inhibitor of p53-dependent transcription and a growth-promoting pro-oncogene [87]. Since Barret’s oesophagus is damaged by reflux of stomach acid into the oesophagus, it is interesting to see if acid can influence AGR2 expression since it is upregulated in this type of disease.

Additionally, high levels of AGR2 mRNA have been found in pancreatic cancer tissue compared to paired normal tissue [143]. Further investigation complement this study where AGR2 was described as a transforming growth factor-β (TGF-β)-responsive gene in human pancreatic cancer cells, whose downregulation is

SMAD4 dependent [76]. In this study also, AGR2 exhibited chaperone activity in which it involves in the maturation of the cancer-associated mucin, MUC1 and that AGR2 is both sufficient and required for MUC1 expression in pancreatic cancer cells [76]. AGR2 also has been implicated in another various type of cancers, among others, gastric cancer cells [140], colorectal cancer [155], lung cancer [142] [156] as well as head and neck cancers [157]. AGR2 implications in the cancer landscape are summarised in Table 1.4.

Cancer Type	AGR2 Implication	Reference
Breast	Cell proliferation and migration	Li *et al.,* 2015 [158]
	Overexpression	Fletcher *et al.*, 2003[58]
	Associated with ERBB2 tumours	Duran *et al.*, 2008 [159]
	Poor prognosis	Barraclough *et al.*, 2009; Hrstka *et al.*, 2010 [127, 130]
	Metastasis	Fletcher *et al.*, 2003; Liu *et al*., 2005; Barraclough *et al*., 2009 [58, 129, 130]
Ovarian (mucinous)	Overexpression	Park *et al.*, 2011 [137]
	Overexpression	Gray *et al.*, 2012 [136]
	High plasma concentration	Edgell *et al.*, 2010 [135]
Ovarian (other)	Poor prognosis	Darb-Esfahani *et al.*, 2012 [149]
	Overexpression	Armes *et al.*, 2013 [148]
Prostate	Overexpression in serum	Kani et al., 2013 [151]
	Overexpression	Kristiansen *et al.*, 2005; Bu *et al.*, 2011 [133, 150]
	Metastasis	Kani *et al.*, 2013; Zhang *et al.*, [151, 160]
	Cellular senescence	Hu *et al.*, 2012 [161]
	Poor prognosis	Zhang *et al.*, 2007 [134]
Oesophageal	Overexpression	Pohler *et al.*, 2004 [87]
Pancreatic	Overexpression in serum	Barry *et al.*, 2012; Ramachandran., 2008 [143, 162]
	Secreted into pre-malignant neoplastic juice	Chen *et al.*, 2010; Makawita *et al.*, 2011 [163, 164]
	Promotes pancreatic intraepithelial neoplasia	Norris *et al.*, 2012 [76]
Lung	Overexpression	Fritzsche *et al.*, 2007 Chung *et al.*, 2011 [142, 156]
Gastric	Metastasis	Lee do *et al.*, 2011 [140]
Colorectal	Poor prognosis	Valladares-Ayerbes *et al.*, 2012 [155]
Head and neck	Metastasis	Sweeny *et al.*, 2012 [157]

Table 1.4 AGR2 implication in various type of cancers.

1.3.6.2. AGR2 implication in metastasis and excretory system

The AGR2 metastatic role is intensively studied in many cancer types including breast carcinomas [58, 129, 130], prostate adenocarcinomas [151], head and neck squamous cell carcinomas [157] or gastrointestinal cancers [140]. A key nature of AGR2 highlighted in many studies is, not only the effect of AGR2 in growth promotion, but also the detection of AGR2 in excretory systems such as blood serum and urine, suggesting that AGR2 may act in a paracrine or autocrine manner [165]. AGR2 was identified as a candidate marker for metastasis in a gene expression screen of cancer cell lines [166]. Further interactomic studies have demonstrated that AGR2 directly interacts with two transmembrane proteins involved in metastasis-associated GPI-anchored protein cell-cell contact or cell-matrix interactions in C4.4A and alpha-dystroglycan (DAG1) respectively (Table 1.3) [58]. cDNA microarray analysis of ovarian tumour tissue revealed AGR2 overexpression and its overexpression upregulates the expression of genes involved in cell proliferation, invasion and angiogenesis, which are very important for tumour progression and metastasis [137]. Moreover, overexpression of AGR2 in benign rodent mammary cell line transformed the cells to the metastatic phenotype and showed enhanced adhesive properties, suggesting AGR2 significance on the *in vivo* tumour biology [129]. In head and neck squamous cell carcinoma cells model, AGR2 silencing (and also silencing of CD147) reduced cell proliferation, migration and invasion [157]. Thus, these results suggest that AGR2 does not only promote cell migration but can also control the cell adhesion rates of detached cells, which may relate to the metastatic properties. The differences in whether migration or adhesion dominates through an AGR2 signal are likely because of different cancer cell models were used in distinct experiments and that each cell type exploits distinct pathways that mediate cancer survival.

AGR2 also has been shown to be associated with metastatic in prostate cancer. AGR2 overexpressing cells in non-metastatic prostate cancer cells promotes the motility and invasiveness, whereas silencing of AGR2 eliminates this behaviour [160]. While AGR2 promotes migration and invasion of prostate cancer cells as well as other types of cancers, it has the opposite phenotype in prostate cancer, it reduces cell-cycle progression [152]. In this study, overexpression of AGR2 led to cell growth inhibition

and reduce colony formation of prostate cells, conceivably brought by a cell cycle arrest [152]. This suggests that AGR2 may involve in regulating epithelial–mesenchymal transition (EMT) in development and cancer and cell-proliferation effect of overexpressed AGR2 is dependent on cell type. It has been reported that SIP, a transcription factor which induced an invasion phenotype but at the same time attenuates cell growth control EMT. EMT is a process by which epithelial cells lose their cell polarity and adhesion and gain migratory and invasive properties to become mesenchymal cells. EMT was a major programme used in cell differentiation during embryogenesis, but parts of this genetic programme are believed to be reactivated during metastasis to transform malignant epithelial cells into increase motility to travel to the distant site [167, 168]. Possibly, AGR2 is suppressed during EMT but reactivated again once metastasized to distant organs.

The cellular mechanism by which AGR2 promotes proliferation are poorly understood, but there are several hints from AGR2 studies on the role it plays in cell signalling networks. One pathway that AGR2 has been implicated in is the EGFR pathway. Dong *et al.* demonstrated that AGR2 induced expression of AREG, a growth promoting EGFR ligand [169] which functionally associated AGR2 and AREG and supported a significant role for AGR2 in lung adenocarcinomas and the regulation of cell growth. AREG induced by AGR2 expression may stimulate the EGFR signalling pathway and may be responsible for the increased cell proliferation and anchorage-independent growth observed in transformed cells [62, 143]. It was also shown that the induction was specific for AREG, as there was no detection of other known EGFR ligands and that activation resulted in increased phosphorylation of both EGFR itself and AKT downstream of the receptor. AREG expression has also been shown to be regulated by the Hippo pathway, which involve in the regulation of cell growth and apoptosis, and functions in dictating organ size [170]. Suppression of the Hippo pathway results in YAP dephosphorylation followed by transport to the nucleus, where it inhibits apoptosis and promotes cell division. Nuclear YAP protein is associated with neoplasia and has been observed in lung, colon, ovarian and breast adenocarcinoma [169].

Further, AGR2 was differentially expressed in metastatic gastric cancer cells, compared to non-metastatic cells using proteomics approach [140]. The role of AGR2 in metastasis is of significant interest, as successful metastases require activation of genes that promote cell survival in environments of stress, therefore the necessity to characterise this metastatic marker proved to be clinical significance.

1.3.6.3. AGR2 and cancer drug resistance

The clinical relevance of AGR2 is strengthened by the finding that AGR2 showed resistance to major cancer drug therapies (as summarised in Table 1.5). For example, AGR2 is one of the genes found to be mediating the resistance of breast cancer drug, Tamoxifen. Tamoxifen is an anti-oestrogen that can be used successfully to treat human breast cancers but the treatment has become a problem due to intrinsic or acquired to the drug. This led to the hypothesis of some genes paradoxically act as agonists of Tamoxifen and might mediate this resistance. As such, identification of oestrogen responsive genes induced by Tamoxifen was performed to identify key players in the field of drug resistance [171]. AGR2 expression is elevated in response to tamoxifen treatment [172]. In cancer patients treated with the oestrogen suppressor Letrazole, the AGR2 gene was one of the top downregulated genes in biopsies from patients' post-treatment that respond well to the drug [127]. This observation suggests that resistance to anti-oestrogens might be due to failure to suppress AGR2. Another study using label-free quantitative proteomics screen, PaCIFIC to search for dominant proteins induced by Tamoxifen demonstrated that AGR2 was the second most upregulated protein [173]. The study also showed that AGR2 gene can promote resistance to cell death induced by the common DNA-damaging agent doxorubicin and AGR2 grown in xenograft system confer resistance to cisplatin [173]. It was also demonstrated that induction of AGR2 was a direct effect of ERα activation using chromatin immunoprecipitation and reporter assay in the presence of oestradiol or Tamoxifen [127]. Later, it was found that AGR2 involve in the phosphoinositidine-dependent protein kinase 1-AKT (PDPK1-AKT) signalling pathway in response to tamoxifen treatment [174].

Additionaly, AGR2 has been associated in an anti-oestrogen resistance comparing fulvestrant-sensitive and –resistant T47D breast cell lines in proteomic and genomic studies [175]. Conversely, AGR2 expression is reduced in docetaxel-resistant hormone-refractory prostate cancer using proteomic screens [176].

Tumour cells may overexpress AGR2 to promote cell survival by allowing the cell to withstand excess protein production associated with transformed cells. Major histocompatibility complex-1 (MHC-1) molecules, important for tumour surveillance, are potential targets of AGR2 as other PDIs have been shown to play a role in their folding and antigen loading[177]. Most cancer cells show an increased expression of proteins of the secretory machinery, including ER chaperones that contribute to the ability of cancer cells to confer resistance against a wide range of therapies. For example, it is known that the simple overexpression of the central chaperone BiP has been shown to be protective against a wide range of chemotherapeutic agents [178].Since AGR2 can exist intracellularly and also extracellularly, AGR2 may involve in the secretory machinery and thus upregulation of AGR2 in cancer can lead to overexpression of other the secretory machinery. Thus, understanding the molecular details of why certain cancer drug selectively induces or reduce AGR2 expression might help to identify novel drug targets to circumvent these problems.

DRUG	INVOLVEMENT OF AGR2	REFERENCE
TAMOXIFEN	Induced by tamoxifen, overexpressed in tamoxifen-resistant tumours	Hrstka *et al.*, 2010[127]; Hengel *et al.* 2011 [173]
CISPLATIN	Mediates resistance in xenograft models	Hengel et al 2011[173]; Gray *et al.* 2012 [136]
DOXORUBICIN	Overexpressed in doxorubicin-resistant tumours	Hrstka *et al.*, 2010 [127]; Hengel *et al.* 2011 [173]
DOCETAXEL	Reduced expression in docetaxel-resistance tumours	Zhao *et al.* 2009 [176]
FULVESTRANT	AGR2 and Cathepsin D mRNA upregulated in resistant cell line	Huber *et al.* 2004 [175]

Table 1.5. Studies underlining AGR2 in drug resistance.

1.3.7. AGR2 in the early secretory pathway and extracellular gain-of-function.

Enhanced AGR2 expression is observed in many cancers as described previously and this can be termed to intracellular AGR2 (iAGR2) expression. iAGR2 overexpression could represent a mechanistic intermediate between the endoplasmic reticulum and tumour development. For example, Higa et. al showed that iAGR2 expression could enhance ER protein homoeostasis and or proteostasis to increase the capacity of tumour cells to handle the demands for protein production and secretion thereby contributing to the aggressiveness of cancer [75]. Although the iAGR2-mediated ER proteostasis control model is interesting, it was also observed that in cancer, AGR localises in a different part of the cells and is secreted (Figure 1.11 [94]). AGR2 can escape ER retrieval machinery and can be found in the cytoplasm as well as in the extracellular environment probably due to possession of the non-optimized ER retention site (KTEL) at its C-terminal [58]. AGR2 was found to be secreted during the development of pancreatic cancer [143] and was detected in the urine of prostate cancer patients [152] as well as in the blood of ovarian cancer patients [135]. These non-endoplasmic reticulum locations define broader functions of that may be based on different redox potential based on the localisation [179]and thereby opening other openings for its role in tumour microenvironment quality control and tumour development.

More interestingly, a recent study describes the physiological role of extracellular AGR2 (eAGR). Fessart *et al.* used an organoid model derived from human lung epithelial cell to address whether eAGR2 could act as a pro-oncogenic molecule in the extracellular matrix[180]. The study showed that eAGR2 present in the extracellular medium of non-tumoural organoids can convert non-tumoural organoids to tumour organoid and enhanced their growth about ten-fold [180]. Interestingly, eAGR2 is sufficient by itself, to disrupt cell polarisation and to promote the acquisition of invasive and metastatic properties and this is independent of its thioredoxin-like domain (CXXS) and of its ER-retention domain (KTEL). This study highlights a gain-of-function of AGR2 in extracellular milieu of the tumour microenvironment. Therefore, it can be postulated that there is a dual role of AGR2 in

cancer development. First, there is an intracellular AGR2 whose function previously described as a catalyst in the ER-homeostasis to cope with cancer cells secretory demands an. Second, an extracellular AGR2 that involves in pro-oncogenic signalling in epithelial tumorigenesis. Altogether these findings underscored AGR2 as an appealing target for novel therapeutic strategies against cancer.

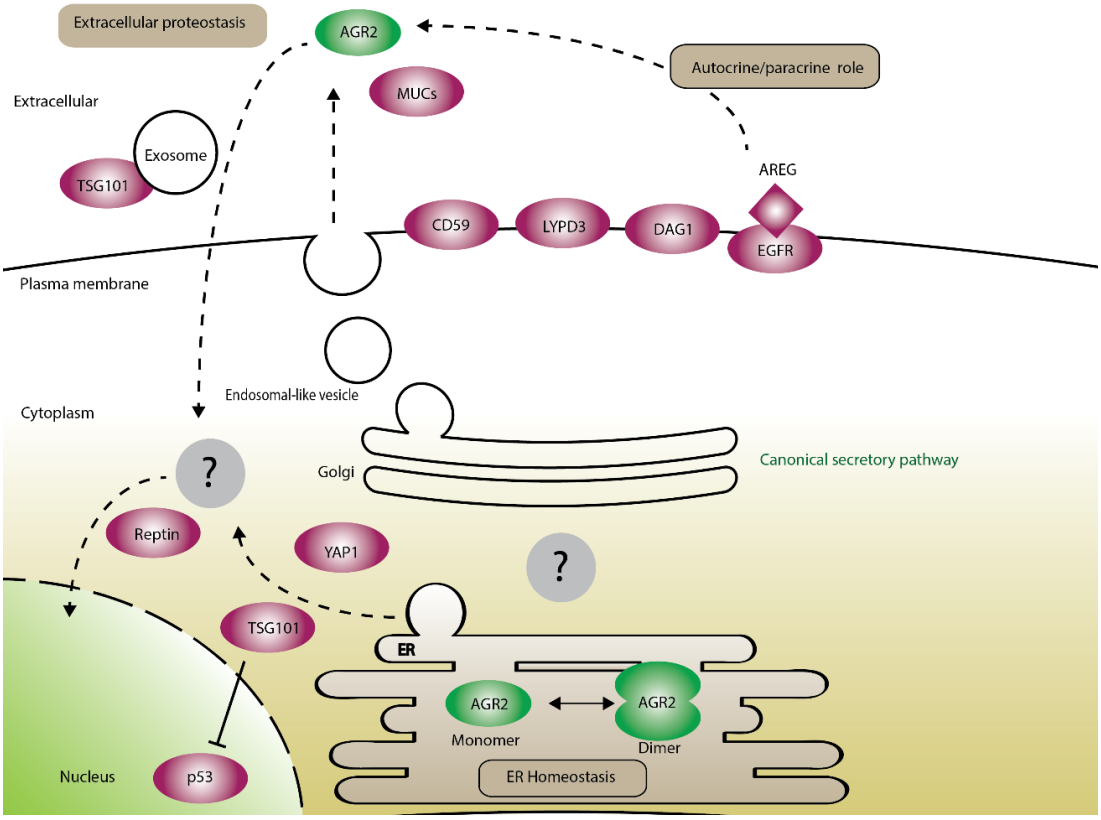


Figure 1.11 The intracellular and extracellular role of AGR2. (From [116]) Although AGR2 normally reside in the ER (ER homoeostasis), studies have shown that AGR2 can be identified outside the ER and secreted in extracellular space (autocrine/paracrine role or proteostasis). There are very few AGR2 "binding proteins" identified; one of which can bind AGR2 cytosolically in a complex with the AAA+ protein Reptin where the dimeric AGR2 shows enhanced affinity. Another AGR2 client proteins includes the exosomal and ubiquitin ligase family member TSG101 that can play a direct role in suppressing p53 function.

1.4. Aims

The main aim of this study was to expand the AGR2 interactome and to understand the functional consequences of some of these interactions. Chapter 3 discussed the use of linear peptide motif that specifically binds to AGR2 in which; i) the peptide binding site was mapped using hydrogen-deuterium mass spectrometry; and ii) the consensus peptide-binding motif was developed to search for human proteins that harbour this motif as potential AGR2 client proteins. When the potential AGR2 client proteins were grouped based on their subcellular localisations, transmembrane proteins were dominant class of proteins. This leads to another aim in the project which is to study AGR2 involvement in secretory pathway since AGR2 has the propensity to bind to these secretory proteins and AGR2 itself can be secreted for extracellular gain-of-function. The key AGR2 client protein represented by a membrane-bound protein EpCAM was validated in-depth using in-vitro biochemical approaches and mass spectrometric analysis. Chapter 4 was built around the key AGR2 client protein (EpCAM) validated from Chapter 3 where the chapters aim; i) to further understand AGR2-EpCAM interaction using cell-based assays; ii) to understand how AGR2 can regulate EpCAM expression using molecular and cellular biology techniques; iii) and to examine the relevance of its expression in tumour clinical samples. Chapter 5 discussed the attempts to deplete AGR2 and to control AGR2 overexpression in tumour cells as means to study its function which is then used to screen for global protein changes using quantitative utilised quantitative proteomics. This also serves as the second method to identify AGR2 client proteins and to support the working paradigms that AGR2 largely binds to membrane-bound secretory proteins. Finally, synthetic tools were utilised to further understand the function, regulation and 'druggability' of AGR2 in Chapter 6. These synthetic tools include the i) isolation of high-affinity AGR2-targeting monoclonal antibodies in the format of the scFv fragment, ii) engineering synthetic mini-protein containing copies of wild-type and mutated AGR2 linear peptide motif and iii) engineering synthetic membrane protein model that bind to AGR2. These synthetic tools were modified to express in mammalian cells and assessed how their expression is regulated by AGR2 to gain information regarding AGR2 signalling and to highlight potentially 'druggable' stage in the oncogenic secretory pathway.

CHAPTER 2: Materials and Methods

2.1. Chemicals

All chemicals and solvents were obtained from commercial sources and of high purity and HPLC spectral grade unless otherwise stated.

2.2. Microbiological techniques

All procedures using bacteria were carried out under aseptic techniques.

2.2.1. Bacterial growth culture

Small DNA preparation (miniprep) extraction from bacterial cell culture 5 ml Luria-Bertani (LB) medium (Table 2.1) containing the appropriate antibiotic (Table 2.2) was inoculated with a single bacterial colony from a freshly streaked selective plate or from a glycerol stock store in -80^{0}C and incubated overnight at 37^{0}C with shaking at 200 rpm. For larger scale DNA preparation (maxiprep) 1-5 ml LB medium containing the appropriate antibiotic was inoculated with a bacterial colony grown on agar or from a glycerol stock and incubated for 8 hours at 200 rpm. The starter culture is diluted 1/500 to 1/1000 into a larger volume 200 ml LB containing appropriate antibiotic before incubation at 37^{0}C with shaking at 200 rpm and grown to saturation (16 hours). The flasks used were of at least four times greater than the volume of medium for proper aeration. Glycerol stocks were prepared as appropriate by transferring 0.8 ml of saturated culture to 0.2ml of sterile 80% (v/v) glycerol using a sterile cryovial (Nunc) before storing at -80^{0}C freezer.

LB Broth	LB plates
1% (w/v) Tryptone	1% (w/v) Tryptone
0.5% (w/v) Yeast extract	0.5% (w/v) Yeast extract
1% (w/v) NaCl	1% (w/v) NaCl
	1.5 % (w/v) Agar

Table 2.1. Bacterial culture medium and plates

Antibiotics	Stock concentration	Storage	Working dilutions
Ampicillin (Sigma)	100 mg/ml	-20^{0}C	100µg/ml
Kanamycin (Sigma)	50 mg/ml	-20^{0}C	50µg/ml

Table 2.2. Antibiotics selection for bacterial growth

2.2.2. Preparation of competent cells

Bacteria cells (DH5α, BL21 DE3 or BL21 AI) from glycerol stocks were inoculated into 5 mL of LB medium and incubated overnight in an incubator-shaker at 37ºC, 200 rpm. LB broth (5 ml) was inoculated with a glycerol stock of DH5α cells and shaken overnight at 37^{0}C. 250 µl of starter culture was used to inoculate 100 ml of LB broth in a 250 ml flask and shaken at 37^{0}C until the culture reached an OD_{600nm} of 0.4. After this step, all procedures were done on ice or in cold room. Cells were pelleted by centrifuging at 4000rpm (Sorvall Legend X1R) for 15 minutes at 4^{0}C, and the pellet was gently resuspended in 32 ml ice-cold competent buffer I (Table 2.3) and left on ice for 10 minutes. The cells were centrifuged as above, and the cell pellet was gently resuspended in 4 mL of ice-cold competent buffer II (Table 2.3) and quickly aliquoted (50 µL) into pre-chilled sterile microcentrifuge tubes. The aliquots were snap-frozen in liquid nitrogen and stored at -80ºC. Transformation efficiency of each batch was checked afterwards by transformation of a competent cell aliquot with plasmid DNA using the heat shock method (see below).

Competent Buffer I	Competent Buffer II
100 mM RbCl	10 mM MOPS
40 mM $MgCl_2.6H_2O$	10 mM RbCl
60 mM CH_3COOK	75m M $CaCl_2.2H_2O$
100 mM $CaCl_2.2H_2O$	15% (v/v) glycerol
15% (v/v) glycerol	Adjust to pH 6.5 with NaOH and sterilise by filtration using 0.22 µm filter
Adjust to pH 5.8 with CH_3COOH and sterilise by filtration using 0.22 µm filter	

Table 2.3. Buffers for making competent cells

2.2.3. Transformation of competent cells

Transformation was performed using heat-shock method unless otherwise stated. 100 to 200 ng of plasmid DNA or 10 µL of the ligation reaction was added to 50 µL of ice-thawed DH5α *E. Coli* competent cells prepared above (for plasmid propagation purposes or BL21 DE3/AI for protein expression) in a microcentrifuge

tube. The mixture was incubated 30 min on ice before incubating at 42°C for 40 sec. Then the tube was quickly cooled on ice for 2 min. Using aseptic techniques, 450 µL of LB broth was added to the tube. Then bacterial cells were incubated with shaking at 37°C for 1 hour. Between 50µl to 100µl of recovered bacterial cells were plated onto 10-cm LB agar plates with appropriate antibiotic and incubated upside down overnight in 37°C incubator. If bacterial cells were less turbid, the cells were centrifuged at 3000 rpm for 5 min, 450 µL of the supernatant was removed and the pellet was gently resuspended in the remaining LB by pipetting up and down before plating on LB plate to get more cells on the plate. Plates were then kept upside down with parafilm around the edge at 4ºC.

2.3. Molecular biology techniques

Plasmids used in this study are listed below.

Plasmid Name	Description	Reference
FL-EPCAM	Encodes Full-length EPCAM with pCMV6-XL5 backbone vector Amp^R	Origene
pmAGR2	Mature AGR2 (aa 21-175). N-term 6xHis-Tag. Amp^R	Dr Euan Murray
pEHISTEV	N-Term 6xHis-tag. Kan^R	Dr Euan Murray
pmEPEX-CO	Human codon-optimised extracellular part of EpCAM (aa 24-265) cloned into pET151/D-TOPO N-Term 6xHis-tag. Amp^R	Life Technologies

Table 2.4. List of plasmid constructs for bacterial expression used in this study.

Plasmid Name	Description	Reference
pEGFP-N1	Encodes the GFPmut1 variant which contains the double-amino-acid substitution of Phe-64 to Leu and Ser-65 to Thr. Kan^R	Clontech
pmCherry-N1	Encodes mCherry, a mutant fluorescent protein derived from the tetrameric *Discosoma sp*. red fluorescent protein (DsRed), Kan^R	Clonetech
pSF-RSV (OG146)	Rouse sarcoma virus (RSV) promoter upstream of the MCS, contains the pUC high-copy origin of replication. Kan^R	Oxford Genetics
pSF-CMV (OG2)	CMV promoter upstream of the MCS contains the pUC high-copy origin of replication. Amp^R	Oxford Genetics
FL-AGR2-C-Cherry	FL-AGR2 cloned into pcDNA3/GW-Cherry. Amp^R	Brno,Czech Republic
FL-AGR2-C-GFP	FL-AGR2 cloned into pcDNA3/GW-GFP. Amp^R	Brno, Czech Republic
FL-EPCAM-C-GFP	FL-EPCAM cloned into pEGFP-N1. Kan^R	Cloned in-house
FL-EPCAM-C-Cherry	FL-EPCAM cloned into pmCherry-N1. Kan^R	Cloned in-house
pAM1L	Synthetic FL-AGR2 WT	Cloned in-house
pAM1H	L= Cloned into pSF-RSV; H=Cloned into pSF-CMV	
pAM2L	Synthetic FL-AGR2 KDEL (Thr173 to Asp). Enhanced ER localisation.	Cloned in-house
pAM2H	L= Cloned into pSF-RSV; H=Cloned into pSF-CMV	
pAM3L	Synthetic mature AGR2 (Δ aa 1-20 aa signal peptide) and ΔKTEL (Δ aa 172-175). Enhanced AGR2 secretion.	Cloned in-house
pAM3H	L= Cloned into pSF-RSV; H=Cloned into pSF-CMV	
pAM4L	Synthetic, deletion of N-term AGR2 intrinsically region (Δ aa 20-40) and ΔKTEL (Δ aa 172-175). Enhanced dimer.	Cloned in-house
pAM4H	L= Cloned into pSF-RSV; H=Cloned into pSF-CMV	
pAM5L	Synthetic FL-EpCAM WT	Cloned in-house
pAM5H	L= Cloned into pSF-RSV; H=Cloned into pSF-CMV	
pAM6L	Synthetic mature EpCAM (Δ aa1-23 signal peptide)	Cloned in-house
pAM6H	L= Cloned into pSF-RSV; H=Cloned into pSF-CMV	

Plasmid Name	Description	Reference
pAM7	Synthetic human codon-optimised FL-EPCAM cloned into pcDNA3.1(+).Amp$^{R.}$	Life Technologies
pAM8	Synthetic human codon-optimised AGR2 scFv4 antibody fragment with 20 aa AGR2 signal peptide fused at N-terminal, CD20 epitope tag and KDEL motif fused at its C-terminal.	Cloned in-house
pAM9	Synthetic human codon-optimised AGR2 scFv4 antibody fragment with truncated AGR2 signal peptide fused at N-terminal, CD20 epitope tag without KDEL motif.	Cloned in-house
pAM10	Synthetic human codon-optimised CHIP scFv (11F) antibody fragment with 20 aa AGR2 signal peptide fused at N-terminal and KDEL motif fused at its C-terminal.	Cloned in-house
pAM11	synPRO^{wt}. Synthetic human codon-optimised AGR2-binding protein containing 3x PTTIYY motifs with 20 aa AGR2 signal peptide fused at N-terminal, CD20 epitope tag and KDEL motif fused at its C-terminal.	Cloned in-house
pAM12	synPRO^{mut}. Synthetic human codon-optimised AGR2-mutant binding protein containing 3x PKTIGS motifs with 20 aa AGR2 signal peptide fused at N-terminal, CD20 epitope tag and KDEL motif fused at its C-terminal	Cloned in-house
pAM13	synMEM^{wt}. Synthetic human codon-optimised membrane protein containing AGR2-binding protein containing 3x PTTIYY motifs with 20 aa AGR2 signal peptide fused at N-terminal, CD20 epitope tag, EpCAM transmembrane domain (aa266-288) and KDEL motif fused at its C-terminal.	Cloned in-house
pAM14	synMEM^{mut}. Synthetic human codon-optimised membrane protein containing mutant AGR2-binding protein containing 3x PKTIGS motifs with 20 aa AGR2 signal peptide fused at N-terminal, CD20 epitope tag, EpCAM transmembrane domain (aa266-288), and KDEL motif fused at its C-terminal.	Cloned in-house

Table 2.5. List of plasmid constructs for mammalian expression used in this study. All synthetic cDNA was obtained from GeneArt® Gene Synthesis, Life Technologies.

2.3.1. Plasmid DNA Purification and Quantification

Bacterial cultures were prepared as above (2.2.1) where 5 ml and 200 ml of overnight bacterial cultures were used for small (miniprep) and large (maxiprep) scale plasmid DNA purification respectively. For small-scale plasmid DNA purification, QIAprep Spin Miniprep Kit (Qiagen) instructions were followed to extract and purify plasmid DNA. For the larger scale of plasmid DNA purification, HiSpeed Plasmid Maxi Kit (Qiagen) or GeneJET Plasmid Maxiprep Kit (LifeTechnologies) instructions were followed for extraction and purification of plasmid DNA, and DNA was eluted with 0.5 to 1 mL of TE Buffer (Qiagen) and stored at -20°C. DNA concentration and purity was measured with a NanoDrop 2000 spectrophotometer (absorbance at 260 nm).

2.3.2. Agarose Gel Electrophoresis

Agarose gel (1-2% (w/v) ultrapure agarose (Life Technologies) in 1X TAE (40 mM Tris pH 7.5, 2 mM $Na_2EDTA.{}_2H_2O$, 0.114% (v/v) acetic acid) was warmed up in a microwave to dissolve the agarose. If the DNA fragment is less than 500 bp , 1X TBE (890mM Tris, 890mM Boric acid, 20mM EDTA pH8.0) was used. SYBR® Safe DNA gel stain (Life Technologies) was added to the agarose solution at a final concentration of 1X. The agarose solution was then poured in a gel cast and let solidify by cooling at room temperature. DNA samples were mixed with 6X DNA loading buffer (0.25% (w/v) Bromophenol blue, 0.25% (w/v) Xylene cyanol, 15% (v/v) glycerol) at a 5:1 ratio of DNA sample to loading dye. Samples were loaded onto the gel alongside Quick-Load 100 bp DNA ladder (NEB) and/or 1 kb DNA ladder (NEB) and were subsequently run at 100 V until the migration front reached the third quarter of the gel. The gel was then visualised under an ultraviolet lamp imaging system (Syngene Chemi Genius 2 Bio Imaging system, Syngene) with a built-in camera (Fluor Cam 210). For samples requiring DNA extraction from the gel, appropriate bands were excised and recovered using the QIAquick Gel Extraction Kit (Qiagen) according to manufacturer's instructions.

2.3.3. Cloning using restriction enzymes

2.3.3.1. PCR amplification

The DNA insert was amplified using a holding plasmid DNA containing the gene of interest. Directional cloning was used where the DNA insert and vector were digested with two different restriction enzymes to create two different complementary ends, thus ensuring only one possible orientation for the insert to effectively ligate to the vector cut with the same pair of restriction enzymes. When designing the primers, one restriction enzyme site is incorporated at each 5' (forward) and 3' (reverse) ends of the DNA insert sequence on the primer. A list of primers with their restriction sites and which template was used for each PCR are given in (Table 2.6). The primer was ensured to be in frame with any open reading frame or tag on the vector. Generally, the primers were 20-30 bp long with melting temperature (Tm) of 65°C and 75°C and GC content between 40-60%.

Vector	Target	Primer Names	Sequence 5'-3'
pEHISTEV	Extracellular EPCAM $_{24-265}$	EpCAM-Ex-F	GGGGCCATGGCTCAGGAAGAATGTGTC
		EpCAM-Ex-R	CCCCCTCGAGTTATTTTAGACCCTGCAT
pmCherry-N1, pEGFP-N1	Full Length EPCAM	FL-EpCAM-F	GCT CTC GAG ATG GCG CCC CCG CAG GTC CTC
		FL-EpCAM-R	CGACCGGTGCATTGAGTTCCCTATGCATCTC

Table 2.6 Primers for DNA amplification

PCR reactions were set up on ice in nuclease-free microtubes according to (Table 2.7). PCR reactions were performed in an Agilent SureCycler 8800 or DNA Engine Dyad™ machine. PFU® Turbo DNA polymerase was used for amplification because it is a high fidelity polymerase and possesses a 3' to 5' exonuclease proof-reading activity that enables the polymerase to correct nucleotide-misincorporation errors. This means

that *Pfu* DNA Polymerase-generated PCR fragments will have fewer errors than *Taq*-generated PCR inserts. Three µl of PCR products mixed with DNA loading buffer was electrophoresed on 1% agarose gel or 2% if the DNA size is lower than 500 bp to check for specific amplification product. Where appropriate the PCR products were cleaned using QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's protocol.

Component	**Final Concentration**
2x Pfu Turbo Master Mix (Rovalab)	1x
10 µM Forward primer	0.5 µM
10 µM Reverse primer	0.5 µM
DNA Template	100ng
	Total volume 20 µl

Table 2.7. PCR reactions

PCR cycles protocol		
Step 1	95°C	2 min
Step 2	95°C	30 sec
Step 3	58°C	30 sec
Step 4	72°C	1 min 30 sec
Step 5	Repeat step 2-4 30 times	
Step 6	72°C	10 min
Step 7	4°C	∞

Table 2.8. PCR cycles protocol for conventional cloning

2.3.3.2. Restriction digestions

PCR products, which have been cleaned up, by either QIAquick PCR Purification Kit or QIAquick Gel Extraction Kit (Qiagen) were double digested using restriction enzymes (NEB) and their compatible buffers according to the manufacturer's guideline. The vector into which the gene was to be inserted into was digested in parallel. The digestion was set up as in table and incubated 2-3 hours at 37^0C water bath.

Component	Final Concentration
Insert/vector	0.5ug - 2µg
NEBBuffer	1x
Restriction enzyme 1	10 units (1µl)
Restriction enzyme 2	10 units (1µl)
	Total volume 20 µl

Table 2.9. DNA Double digestions reaction

The digested products were then mixed with DNA loading buffer and electrophoresed onto 1-2% agarose gel, and single bands corresponding to the double-digested vector and double-digested insert were cut out under ultraviolet light. DNA was extracted using the QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's instructions.

2.3.3.3. Ligation

Ligation of the double-digested gel-extracted insert and vector was carried out using T4 DNA ligase (NEB) according to the manufacturer's instructions. The amount of insert to be added was calculated using the formula:

$$\text{ng of insert} = \frac{\text{ng of vector x kb size of insert}}{\text{kb size of vector}} \times \text{molar ratio of } \frac{\text{insert}}{\text{vector}}$$

Different molar ratios of vector: insert DNA as 1:1 to 1:4 were attempted to increase the chance of successful ligations. Ligation reactions were set up on ice as shown in Table 2.10 and were incubated 3 hours in 37^0C water bath or overnight at room temperature.

Component	Final Concentration
Vector	100 ng
Insert	Calculated molar ratio
10x Buffer (NEB)	1x
T4 DNA ligase (NEB)	10 units (1µl)
	Total volume 20 µl

Table 2.10. DNA ligation reaction

Ten µL were transformed into competent DH5α cells as described before and plated out on LB-Agar plates containing appropriate antibiotic. Single colonies were picked for plasmid DNA isolation. Glycerol stocks were made from the bacterial culture if positive clones are identified later. To verify the presence of the insert in the selected clones, the purified DNA clones were either i) PCR amplified with the same primers used for cloning to see if the insert is successfully ligated or ii) double digested with the same restriction enzymes used for cloning to see if the insert band is visible on agarose gel. Positive clones will be further confirmed by DNA sequencing (2.3.5) to ensure the insert has no mutation and in-frame with the open reading frame or tag.

2.3.4. Site-directed mutagenesis

Mutants of the wild type AGR2 and wild-type EPCAM were made based on the protocol described in QuickChange site-directed mutagenesis (Stratagene). This method is based on a linear PCR amplification whereby mutation-containing primers are incorporated but not copied using high-fidelity Pfu DNA polymerase. The desired single point mutation was designed to be in the middle of the primers as shown in Table 2.11.

Primer Names	Template	Sequence 5'-3'
AGR2-D132A-F	pmAGR2	CC AGG ATT ATG TTT GTT *GCC* CCA TCT CTG ACA GTT AG
AGR2-D132A-R		CT AAC TGT CAG AGA TGG *GGC* AAC AAA CAT AAT CCT GG
AGR2-P133A-F	pmAGR2	C AGG ATT ATG TTT GTT GAC *GCA* TCT CTG ACA GTT AGA GC
AGR2-P133A-R		GC TCT AAC TGT CAG AGA *TGC* GTC AAC AAA CAT AAT CCT G
AGR2-S134A-F	pmAGR2	ATT ATG TTT GTT GAC CCA *GCT* CTG ACA GTT AGA GCC G
AGR2-S134A-R		C GGC TCT AAC TGT CAG *AGC* TGG GTC AAC AAA CAT AAT

Primer Names	Template	Sequence 5'-3'
AGR2-C81S-F	pmAGR2 FL-AGR2-C-Cherry	T CAT CAC TTG GAT GAG *AGC* CCA CAC AGT CAA GC
AGR2-C81S-R		GC TTG ACT GTG TGG*GCT* CTC ATC CAA GTG ATG A
FLEPCAM_Y251A_F	FL-EpCAM-C-GFP	G GAT CTG GAT CCT GGT CAA ACT TTA ATT TAT *GCT* GTT GAT GAA AAA GCA CCT
FLEPCAM_Y251A_R		AGG TGC TTT TTC ATC AAC *AGC* ATA AAT TAA AGT TTG ACC AGG ATC CAG ATC C
FLEPCAMCO_Y251A_F	pmEPEX-CO	CCAGACCCTGATCTAC*GCC*GTGGACGAGAAGGC C
FLEPCAMCO_Y251A_R		GGCCTTCTCGTCCAC*GGC*GTAGATCAGGGTCTGG

Table 2.11. Site-directed mutagenesis primers

The reactions were setup using 2x Pfu Turbo Master Mix (Rovalab) as in Table 2.12 and cycling conditions are described in Table 2.13. All constructs were confirmed and analysed by DNA sequencing.

Component	Final Concentration
2x Pfu Turbo Master Mix (Rovalab)	1x
10 µM Forward primer	0.5 µM
10 µM Reverse primer	0.5 µM
DNA template	100-150 ng
	Total volume 20 µl

Table 2.12 PCR reactions for site-directed mutagenesis

Step 1	95°C	2 min
Step 2	95°C	20 sec
Step 3	58°C	40 sec
Step 4	72°C	7 min
Step 5	Repeat step 2-4 18 times	
Step 6	72°C	10 min
Step 7	4°C	∞

Table 2.13 PCR cycles protocol for site-directed mutagenesis

2.3.5. DNA sequencing

In a small microcentrifuge tube, 10 µl of the sample and 10 µl of primers were prepared as in Table 2.14. The samples were sent to SourceBioscience Scotland for Sanger sequencing. The facility's stock primers were used if available. Where the stock primers are not available, either the forward or the reverse primers used for cloning was used.

Type of samples	Concentration
Plasmid DNA	100 ng/µl
PCR product	1 ng/µl per 100bp
Primers	3.2 pmol/µl

Table 2.14. DNA and primer concentrations for DNA sequencing

The DNA sequence analysis was performed using Chromas v1.45 or SnapGene v3.0 software and aligned using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/).

2.4. Cell culture and handling

All cell culture work was performed in a laminar flow hood using standard aseptic tissue culture procedures. Only molecular biology grade reagents and plasticware were used, otherwise sterilising by filtering or autoclaving was carried out.

2.4.1. Cell lines and maintenance

Cell lines were grown in humidified incubators equipped with CO_2 gas cylinders. Media was supplemented with fetal bovine serum with or without the presence of antibiotics as described in Table 2.15. Cells were maintained in 10-cm tissue culture dishes (Greiner CELLSTAR®, Sigma) and were subcultured 1 to 3 times per week. The medium was removed and cells were washed using 10 mL of sterile 1X phosphate-buffered saline (PBS). Cells were then trypsinised using 1 to 2 mL of 1X Trypsin-EDTA 0.5% (Gibco, Life Technologies) warmed up to 37°C in a water bath beforehand, and cells were returned to the incubator until they detached (3 to 5 min). Warmed medium supplemented described below was added up to a final volume of 10 mL and cells were then seeded in new tissue culture dishes at a dilution of 1:5 to 1:20.

When cells need to be counted before sub-culturing, trypsinised cells were mixed 1:1 with Trypan Blue Stain 0.4% (Labtech) to stain for dead cells and loaded into Luna cell counting slides (Labtech) and counted using Luna Automated BF Cell Counter (Labtech).

Cell Lines	Description	Growth Conditions	Media
OE19	Adenocarcinoma of gastric cardia/oesophageal gastric junction	5% CO_2, 37^0C	RPMI 1640, 10% FBS, 1% Penicillin-Streptomycin when not transfecting
OE33	Adenocarcinoma of the lower oesophagus (Barrett′s metaplasia)	5% CO_2, 37^0C	RPMI 1640 10% FBS, 1% Penicillin-Streptomycin when not transfecting
FLO-1	Primary distal oesophageal adenocarcinoma	5% CO_2, 37^0C	DMEM, 10% FBS, 1% Penicillin-Streptomycin when not transfecting
MCF-7	Ductal carcinoma of the breast	5% CO_2, 37^0C	DMEM, 10% FBS, 1% Penicillin-Streptomycin when not transfecting

Table 2.15 Cell lines and growth conditions

1.4.1. Cell storage

Cell lines were grown to 90% confluence in 10cm dishes and trypsinised as described above (2.4.1)before diluting the trypsinised cells in a final volume of 10ml. Cells were pelleted by centrifuging at 1000rpm (Sorvall Legend X1R) for 5 minutes in sterile 15ml falcon tubes. The cell pellet was resuspended in freezing medium (50% FBS, 20% DMSO and 30% serum-free media). Cells were aliquoted in 1ml aliquots into cryotubes (Nunc) and placed in gradient freezing containers (Nalgene) for 24 hours at -80°C, before transferring to liquid nitrogen containers for long-term storage.

2.4.2. Cell recovery

Cells in cryotubes stored in liquid nitrogen were placed into a water bath at 37°C by swirling the tube until a tiny amount of ice remaining before being transferred into

10 cm dish containing 9 ml of complete media (Table 2.15) and incubated into 37°C. Media was replaced the following day to remove DMSO and cells were grown and maintained as described above (2.4.1).

2.4.3. Transient Transfection of Plasmid DNA

Cells were grown as above to 80-90% confluency before plating according to cell type and experimental conditions. Per 6 well-plate, cells were normally split 1:4 to 1:10 (MCF-7 1:10, OE33-1:5 OE19, 1:5 FLO-1 1:10) or counted one day before the transfection and incubated into 37C to achieve between 40-70% cells confluency the following day. The following the day, transfection were carried out using Attractene Transfection Reagent (Qiagen) or Lipofectamine 2000 (Thermo Fisher Scientific). For transfection using Attractene, per 6-well plate, 1 µg DNA (only DNA from Maxi-prep) was diluted in a total of 100 µl of serum free media without any antibiotics in a sterile microcentrifuge tube. Five µl of Attractene was added to the tube and mixed before incubating at room temperature for 15 minutes for complex formation to take place. For transfection using Lipofectamine 2000, cells were split prior with normal media without antibiotics. Per 6-well plate, 2 µg DNA was diluted in a total of 250 µl OPTI-MEM® reduced-serum medium (Thermo Fisher Scientific) in a sterile microcentrifuge tube. Five µl of Attractene was added to the tube and incubated 5 minutes before mixing with diluted DNA in OPTI-MEM®. Then the mixture was incubated for 20 minutes at room temperature for complex formation to take place. Media in wells split the day before was replaced with 2ml fresh media. The transfection complexes were then added dropwise to the wells and swirled before incubating at 37°C for 24-48 hours or as stated. For experiment using other than 6 well plate, manufacturer's protocol was followed.

2.4.4. Transient transfection of short interfering RNA (siRNA)

MCF7 cells were counted and plated at 200,000 cells per 6-well plate on the day before the transfection. The cells were transfected with siRNA, control siRNA, or mock transfected as above using Attractene (2.4.3). The siRNA to AGR2 used was from FlexiTube GeneSolution GS10551 for AGR2 (#1027416, Qiagen) and control AGR2 siRNA used was AllStars Negative Control siRNA (#SI03650318 , Qiagen).

For siRNA to AGR2 in MCF-7, 60nM final concentration of siRNA was used and incubated for 72h at 37°C before analysing by SDS-PAGE and western blot.

2.4.5. Harvesting cells

Cell culture dishes were placed on ice to prevent proteolysis. Cell culture medium was discarded and cells were gently and quickly washed with 2x ice-cold PBS (1 mL for a well of a 6-well dish, 2 mL for a 6-cm dish or 5 mL for a 10-cm dish). In the final wash, the cells were scraped on ice in a small volume of fresh PBS (100 µL for a well of a 6-well dish, 200 µL for a 6-cm dish or 500 µL for a 10-cm dish) then transferred to a microcentrifuge tube before centrifuging cells at 3000 rpm for 5 min at 4°C. The supernatant was discarded and the cell pellet was either snap-frozen in liquid nitrogen and stored at -80°C, or immediately lysed using the appropriate buffer (see below 2.4.6). In some cases detailed when relevant, cells were directly scraped into lysis buffer.

2.4.6. Cell lysis

2.4.6.1. Denaturing Lysis

Frozen cell pellets were lysed in approximately 4 volumes (with respect to the size of the cell pellet) of denaturing lysis buffer (Table 2.16) by pipetting up and down 10 times. Lysates were then incubated on ice for 30 min, followed by centrifugation at 13,000 g for 15 min at 4°C. The supernatant (lysate) was transferred to a new tube and then snap-frozen and stored at -80°C before protein concentration determination.

Denaturing lysis Buffer	Non-denaturing lysis buffer
7M Urea	0.1 mM EDTA
0.1M DTT	2mM DTT
0.1% Triton X-100	0.1% NP40 (IGEPAL)
25mM NaCl	150mM NaCl
20 mM HEPES-KOH pH 7.6	50 mM HEPES-KOH pH 7.6
1mM Benzamidine	1mM Benzamidine
1X Protease inhibitor mix (Table 2.17)	1X Protease inhibitor mix (Table 2.17)

Table 2.16 Cell lysis buffers

10X Protease inhibitor mix
200 µg/ml Leupeptin
10 µg/ml Aprotinin
20 µg/ml Pepstatin
10 mM Benzamidine
100 µg/ml Soybean trypsin inhibitor
20 mM Pefabloc
10mM EDTA

Table 2.17 Protease inhibitor mix

2.4.6.2. Non-Denaturing Lysis

Frozen cell pellets were lysed in approximately 4 volumes (with respect to the size of the cell pellet) of non-denaturing lysis buffer (Table 2.16) were added to the harvested cell pellet. Alternatively, adherent cells were scraped directly into lysis buffer. Samples were vortexed briefly, incubated on ice for 20 min, and then centrifuged at 13,000 g for 15 min at 4ºC. The supernatant (lysate) was transferred to a fresh tube, snap-frozen in liquid nitrogen and stored at -80º.

2.5. Biochemical techniques

2.5.1. Protein quantification

Protein concentration was quantified using the Bradford assay. Bradford reagent was obtained from Bio-Rad Protein assay dye reagent 5x (Bio-Rad; #5000006). The Bradford reagent was diluted to 1x with water prior to use. Bovine serum albumin (BSA; #A9647, Sigma) was used as protein standards from 0 to 4 mg/ml. Two hundred µl of Bradford reagent was added into a 96-well plate. Then 1 µl of protein lysate was added to the Bradford reagent. The plate was shaken for 5 min at room temperature and the OD_{595nm} was measured using VICTOR^3 Multilabel Plate Reader (Perkin Elmer) with WorkOut v2.0 (Dazdaq) software. Readings were automatically converted to concentrations in mg/mL and concentrations were determined in a comparison to linear standard curve taken from the BSA standard values.

2.5.2. SDS-PAGE

Polyacrylamide gels were prepared in the laboratory. The gels were cast using Mini-PROTEAN 3 Cell apparatus (Bio-Rad). Stacking and separating gels were

prepared as described in (Table 2.18) with either 1.0mm-gap glass plates or 1.5mm gap glass plates dependent on the volume required for loading and allowed to polymerise at room temperature. Upon addition of stacking gel, 10-well or 15 well combs were used depending on assay and number of samples. For analysis of proteins of high molecular weight 10% polyacrylamide gels were used, 12-15% gels were used for lower molecular weight proteins.

10% Separating gel	**12% Separating gel**	**15% Separating gel**	**Stacking gel**
10% Acrylamide (Protogel 30%)	12% acrylamide (Protogel 30%)	12% acrylamide (Protogel 30%)	5% acrylamide (Protogel 30%)
375mM Tris pH8.8	375mM Tris pH8.8	375mM Tris pH8.8	125mM Tris pH8.8
0.1% SDS (w/v)	0.1% SDS (w/v)	0.1% SDS (w/v)	0.1% SDS (w/v)
0.1% APS (w/v)	0.1% APS (w/v)	0.1% APS (w/v)	0.1% APS (w/v)
0.04% TEMED (w/v)	0.04% TEMED (w/v)	0.04% TEMED (w/v)	0.1% TEMED (w/v)

Table 2.18. Separating and stacking gels for SDS-PAGE

Generally 10-50µg (to adjust volume/quantity) protein lysate and 2 µg purified protein was mixed 1:1 with 2X SDS sample buffer (Table 2.19) and boiled at 85°C for 3-5 minutes before loading onto the gels. Five µl of pre-stained protein standards were loaded as a marker (PageRuler™ Plus Prestained Protein Ladder, 10 to 250 kDa, Thermo Fischer), and proteins were separated by electrophoresis at 180-200V using 1x SDS running buffer (Table 2.19) for 45-60 minutes.

2X SDS sample buffer	**1X SDS running buffer**
250mM Tris pH6.8	192 mM Glycine
50% Glycerol	25mM Tris base
5% SDS (w/v)	0.1% SDS (w/v)
Bromophenol blue to the desired colour	
200mM DTT or 10% ß-Mercaptoethanol (added fresh prior to use)	

Table 2.19 SDS sample buffer and running buffer

2.5.3. Western Blot

After separating proteins by SDS-PAGE, proteins were transferred to 0.2 μM pore nitrocellulose blotting membrane (Amersham Protran Premium, GELifesciences) using The Mini Trans-Blot® wet electroblotting systems (Bio-Rad). The transfer was carried out electrophoretically in transfer buffer (Table 2.21) at 300 mA for 90 min or at 20mA overnight. Optionally, Membranes were stained using India ink (Pelican Black) to confirm the proteins were transferred properly. Membranes were washed briefly with PBS-Tween-20 (0.1%) (0.1% PBST) before blocking for 1h at room temperature or overnight at 4°C in blocking buffer (Table 2.21). Primary antibodies (Table 2.20) were diluted in blocking buffer as required and incubated 1 hour at room temperature or overnight at 4°C. Blots were then washed 3x 5 minutes in a shaker before incubation with appropriate secondary antibody diluted in blocking buffer for 1h at room temperature. Blots were washed 3x 5 minutes again in wash buffer. Blots were then visualised by overlaying enhanced chemiluminescence (ECL) reagents (Table 2.22) in a 1:1 ratio for 1 minute. Excess ECL solution was removed using tissue paper and blots were exposed to film (Kodak, Hyperfilm) which was then developed with an X-ray film processor (SRX-101A Film processor , Konica Minolta).

Primary Antibody	Type	Dilution	Product Number/Source	Secondary Antibody
Anti-AGR2*	Rabbit polyclonal	1:1000	K47 (Moravian Biotech)	Swine anti Rabbit 1:2000 (Dako)
Anti-AGR2	Mouse monoclonal	1:1000	AGR3 4.1 (Moravian Biotech)	Rabbit anti-mouse 1:1000 (Dako)
Anti-EpCAM	Rabbit polyclonal	1:1000	HPA026761 (Sigma)	Swine anti Rabbit 1:1000 (Dako)
Anti-EpCAM*	Mouse monoclonal	1:1000	VU-1D9 (Calbiochem)	Rabbit anti-mouse 1:1000 (Dako)
Anti-p53	Mouse monoclonal	1:500	DO-1 (Moravian Biotech)	Rabbit anti-mouse 1:1000 (Dako)
Anti-GFP*	Rabbit polyclonal	1:1000	A-11122 (Thermo Fischer Scientific)	Swine anti-Rabbit 1:2000 (Dako)
Anti-mCherry*	Mouse monoclonal	1:1000	1C51, ab125096 (abcam)	Rabbit anti mouse 1:1000 (Dako)
Anti-β-actin	Mouse monoclonal	1:5000	A5441 (Sigma)	Rabbit anti mouse 1:1000 (Dako)
Anti-CD20*	Mouse monoclonal	1:1000	NCD1.2 (Moravian Biotech)	Rabbit anti mouse 1:1000 (Dako)
Anti-Dermcidin	Mouse monoclonal	1:1000	E-7 (sc-393278, Santa Cruz Biotech	Rabbit anti-mouse 1:1000 (Dako)
Anti-Dermcidin	Rabbit polyclonal	1:500 – 1:1000	#ab175519, Abcam	Swine anti-Rabbit 1:2000 (Dako)

Table 2.20 Primary antibodies for western blot. (*Also used for immunofluorescence as described in 2.8.2)

Transfer Buffer	PBS-Tween20 buffer	Blocking buffer
24 mM Tris	1x PBS	1x TBS
191 mM glycine	0.1% Tween20	0.1% Tween20
20% (v/v) methanol		5% Dried Skimmed Milk

Table 2.21 Western Blot buffers

ECL I	ECL II
2.5 mM Luminol	0.02% (v/v) Hydrogen peroxide
0.4 mM p-Coumaric acid	0.1M Tris pH 8.8
0.1M Tris pH8.8	

Table 2.22 Enhanced chemiluminescence reagents

2.6. Protein Purification

2.6.1. AGR2 purification

The mature form of $AGR2_{21-134}$ was cloned into pDEST17 vector harbouring AGR2 having N-Terminal His-tag and was purified based on His-tag affinity chromatography. *E. coli* strains BL21-AI were transformed with the His-AGR2 construct and were grown aerobically in a 10 ml starter culture overnight at 37 °C at 200 rpm in LB medium supplemented with ampicillin (100 μg/ml). The cultures were then added into 1L LB medium supplemented with ampicillin (100 μg/ml) and 37 °C at 200 rpm. Cultures were induced at OD_{600nm} of 0.6 by the addition of 0.2 % arabinose (w/v) for 3 hours at 37 °C. Cell pellets were collected by centrifugation 10,000 rpm for 10 minutes at 4^0C and resuspended in 10 ml buffer A (Table 2.23)(10ml/litre culture) into 50 ml falcon tube. The cells were then snap freeze using liquid nitrogen and stored at -20°C overnight. The frozen cell pellets were thawed in tap water and resuspended in buffer A with 0.1 mg/ml lysozyme. The resuspended cells were then sonicated 3 × 15 s (MSE Soniprep 150 Plus ultrasonic disintegrator) on ice with 30 sec interval to prevent overheat using a small probe at 10 amplitude. Sonicated cells were then centrifuged at 10,000 rpm for 10 minutes, 4^0C to remove cellular debris. Cell lysates were then mixed with 1 ml Ni-NTA agarose (Qiagen) that was previously washed with 10 ml buffer A wash, for 1 hour in rotating wheel at 4^0C before loading into 10 ml column with 35uM filter pore size (Mobitec) equilibrated with 10ml buffer A. The column was washed with 2x5 ml buffer A followed by 3x5ml with wash buffer

B (Table 2.23). Proteins were eluted with 15x0.5 ml elution buffer C (Table 2.23) in microcentrifuge tubes. AGR2 mutant derivatives His-AGR2 D132A, His-AGR2 P133A and His –AGR2 S134A were purified under the same conditions as above.

Buffers	Components
Buffer A	20mM Tris-HCL pH8, 150 mM NaCl, 10 mM $MgCl_2$, 0.1%NP40, 10% glycerol 20 mM Imidazole
Buffer B	20mM Tris-HCL pH8, 150 mM NaCl, 10 mM $MgCl_2$, 0.1%NP40, 10% glycerol 40 mM Imidazole
Buffer C	20mM Tris-HCL pH8, 150 mM NaCl, 10 mM $MgCl_2$, 0.1%NP40, 10% glycerol 150 mM Imidazole

Table 2.23 His-AGR2 purification buffers

2.6.2. EPCAM purification

Extracellular part of $EpCAM_{24-265}$ was human codon optimised and synthesised into pET151/D-TOPO (GeneArt, LifeTechnologies) which contains N-Terminal His-tag. *E. coli* strains BL21-DE3 were transformed with His-EPCAM construct and were grown aerobically in a 10 ml starter culture overnight at 37 °C at 200 rpm in LB medium supplemented with ampicillin (100 μg/ml). The cultures were then added into 1L LB medium supplemented with 100 μg/ml and 37 °C at 200 rpm. Cultures were induced at OD_{600nm} of 1.0 by the addition of 1mM IPTG for 4 hours at room temperature and 200rpm. Cell pellets were collected by centrifugation 10,000 rpm for 10 minutes at 4^0C. The pellets were resuspended in 20 ml resuspension buffer (25ml/1L culture) (Table 2.24) in 50 ml falcon tube. The pellets were snap-frozen with liquid nitrogen and stored at -20°C overnight. The frozen pellets were thawed in tap water and added 0.5M NaCl, 1mM Benzamidine,1 mM DTT, 0,1% Triton X-100 making up buffer 1 (Table 2.24) to bring the volume to 30 ml. One protease inhibitor tablet dissolved in 1 ml distilled water and 0.1 mg/ml lysozyme was added to the cells. Cells were then sonicated 3 × 15 s on ice with 30 sec interval to prevent overheat using a small probe at 10 amplitude. Solid urea powder (7M) was added directly to the sonicated cells and incubated in the rotating wheel at 4°C until the urea dissolves. Cells were then centrifuged at 10,000 rpm for 5 minutes, 4^0C to remove cellular debris before transferring into a new 50 ml falcon tube containing 0.3 ml Ni-NTA agarose

(Qiagen) that was washed with 50 ml sterile PBS. The beads and the lysate were incubated overnight in a rotating wheel at 4^0C. The Ni-NTA agarose beads were briefly spin down at 1000 rpm for 1 minute, and the unbound lysate was removed. The beads were resuspended in 1 ml wash buffer 2 (Table 2.24) and transferred into a clean microcentrifuge. The beads were washed by incubation in the rotating wheel for 30 min 4^0C. The beads were allowed to sediment at the bottom of the tube by gravity. The buffer was replaced with new 1 ml buffer 2. Unbound proteins were monitored by taking 5 µl of the wash and added to 1X Bradford reagent and look for colour changes. The wash was repeated several times until there were no colour changes on Bradford reagent. To elute the proteins from the beads, the beads were then incubated with 150 µl of elution buffer 3 (Table 2.24) for 45 minutes. The eluate was transferred to a clean microcentrifuge. The elution was repeated 5 times or until the eluate shows no colour changes on Bradford reagent.

Buffers	Components
Resuspension buffer	10% Sucrose (w/v), 50 mM HEPES pH8.0
Buffer 1	10% Sucrose (w/v), 50 mM HEPES pH8.0, 0.5M NaCl, 1mM Benzamidine ,1 mM DTT, 0,1% Triton X-100
Buffer 2	50mM HEPES pH8.0, 0.5M NaCl, 1mM Benzamidine , 1 mM DTT, 2M Urea, 10mM Imidazole.
Buffer 3	50mM HEPES pH8.0, 0.5M NaCl, 1mM Benzamidine , 100mM DTT, 0.3M Imidazole.

Table 2.24 His-EPCAM purification buffers

2.7. In-vitro assays

2.7.1. Protein-Protein interaction using ELISA

2.7.1.1. Peptide ELISA.

96-well polystyrene assay plate (Corning™ Costar™ 96-Well white solid, #07-200-592) was coated with streptavidin by adding 1 µg in 50 µl of water per well and incubated at 37 ° C overnight. The wells were washed 4x with 200 µl 0.1 % PBS-T before adding biotinylated peptides 1-5 µg in 50 µl water per well and incubated 1 hour at room temperature. The wells were washed 6x with 200 µl PBS-T. The wells then were blocked for 1 hour with 200 µl of 3% BSA in PBS-T per well. The blocking

solution was removed and 50 µl of His-AGR2 diluted in 3% BSA in PBST was added to each well and incubated as required for 1 hour at room temperature. The wells were washed 6x with 200 µl PBS-T before probing with 50 µl of primary antibody diluted 1:1000 in 3 % BSA in PBST for 1 hour at room temperature. After washing the wells 6x with 200 µl 0.1% PBST, 50 µl of HRP-conjugated secondary antibody (1:1000 dilution) in 3% BSA in PBS-T was added to each well and incubated for 1 hour at RT. The wells were washed 6x with 200 µl PBS-T before adding 50 µl of enhanced chemiluminescence (ECL) and reading the plate using Fluoroskan Ascent FL as a relative light unit (RLU).

2.7.1.2. Protein binding ELISA.

Protein of interest was coated directly to 96-well polystyrene assay plate using 0.1M $NaHC0_3$ pH 8.6 overnight at 4^0C. Excess free protein in solution was subsequently removed and the wells were washed 4x with 200 µl of 0.1% PBST. The remaining available binding surfaces of the wells were blocked with 3% bovine serum albumin (BSA) in 0.1% PBST. The plate was then incubated with serial dilutions of a second protein of interest (binding partner to the protein coated on the solid phase) to the wells and incubated for 1 hour shaking at room temperature. Remaining free protein in solution was removed and washed 6x with 200 µl of 0.1% PBST, and the portion of the partner protein bound to the immobilised protein was quantified through detection via antibody against the protein in the mobile phase for 1 hour shaking at room temperature. The wells then were washed 6x with 200 µl of 0.1% PBST followed by the addition of HRP-conjugated secondary antibody. The wells of the plate were washed again 6x with 200 µl of 0.1% PBST to remove excess primary antibody and finally detected by adding ECL and reading using Fluoroskan Ascent FL (Labsystems) as relative light unit (RLU).

2.8. Cell biology techniques

2.8.1. Visualisation of fluorescently-tagged proteins

For visualisation of fluorescent-tagged protein, cells were seeded onto glass coverslips in 6 well plate at 50% confluency and incubated at 37 °C incubator overnight. Cells were transiently transfected with DNA constructs encoding GFP or

Cherry-tagged proteins using Lipofectamine 2000 (Thermo Fisher Scientific) as described above (2.4.3) and incubated 37 °C for 24-48 hours. As controls, cells were either untransfected or transfected with GFP and Cherry expressing vector only. Cells on coverslips were then washed three times with sterile PBS before fixation with 4% paraformaldehyde for 15 min at room temperature. Coverslips were washed three times with PBS. Coverslips were mounted with DAPI nuclear stain (Invitrogen) diluted in Fluorescent Mounting Medium (Dako) and viewed using Zeiss Axioplan 2 epifluorescent microscope or Leica TCS SP5 confocal microscope. Epifluorescent images were acquired using a Photometrics Coolsnap HQ2 CCD camera and a Zeiss Zeiss Axioplan II fluorescence microscope with Plan-neofluar objectives (Carl Zeiss, Cambridge, UK), a Mercury Halide fluorescent light source (Exfo Excite 120, Excelitas Technologies) and Chroma #83000 triple band pass filter set (Chroma Technology Corp., Rockingham, VT) with the excitation filters installed in a motorised filter wheel (Ludl Electronic Products, Hawthorne, NY). Image capture was performed using Micromanager[10]. For confocal microscope, images were acquired using a Leica SP5 confocal (Leica Microsystems Milton KeynesUK) using LAS-AF software. The microscope is equipped with a 405nm diode, Argon and, 561 and 648nm laser lines, three Photomultiplier tubes and one HyD GaSP detector. Images were scanned using a 63X or 40x objective.

2.8.2. Immunofluorescence

Cells were seeded onto glass coverslips in 6 well plate at 50% confluency and incubated at 37 °C incubator overnight. Cells were either not transfected (endogenous protein) or transiently transfected with DNA constructs using Lipofectamine 2000 (Thermo Fisher Scientific) as described above (2.4.3) and incubated 37 °C for 24-48 hours. Cells on coverslips were then washed three times with sterile PBS before fixation with 4% paraformaldehyde for 15 min at room temperature. Coverslips were washed three times with PBS and permeabilized with 0.1%Triton X-100 in PBS (v/v) for 10 minutes. Coverslips were washed three times with PBS before incubating with primary antibodies (Table 2.20) in 3% BSA overnight at 4°C. Coverslips were washed three times and secondary antibodies conjugated with fluorophores (Table 2.25) were

[10] https://open-imaging.com/

added and incubated 1 hour at RT. Cells were washed three times with PBS and were mounted with DAPI nuclear stain (Invitrogen) diluted in Fluorescent Mounting Medium (Dako), and viewed using Zeiss Axioplan 2 fluorescent microscope or Leica TCS SP5 confocal microscope as described in 2.8.1.

Secondary antibody name	Dilution
Alexa Fluor® 488 goat anti-mouse	1:1000
Alexa Fluor® 488 donkey anti-rabbit	1:1000
Alexa Fluor® 594 goat anti-mouse	1:1000
Alexa Fluor® 594 donkey anti-rabbit	1:1000

Table 2.25 List of secondary antibodies for immunofluorescence

2.8.3. Proximity ligation assay.

The proximity ligation assay was performed using the Duolink In-situ Detection Reagents Green kit (Sigma-Aldrich). Cells were prepared as described above for fluorescence microscopy. For *in-situ* AGR2-EPCAM interaction, the cells on cover slip were washed 3x with ice cold PBS and incubated overnight with a cocktail of rabbit polyclonal anti-EPCAM (Abcam) and 1:200 and mouse monoclonal anti-AGR3 3.1 1:200 or mouse monoclonal anti-EPCAM (Abcam) 1:250 and rabbit polyclonal anti-AGR2 K47 1:250 diluted in 3% BSA in PBS. Anti-rabbit PLUS probe (Sigma-Aldrich) and anti-mouse MINUS probe (Sigma-Aldrich) were then added to the sections for 1 h at 37 °C. Ligation, amplification, and detection of the probes were carried out according to the manufacturer's protocol. The sections were mounted with DAPI nuclear stain (Invitrogen) diluted in Fluorescent Mounting Medium (Dako) and viewed as described above.

2.9. Single-chain variable fragment (scFv) phage-display

2.9.1. Biopanning procedure

A 96-well polystyrene assay plate (Corning™ Costar™ 96-Well white solid, #07-200-592) was coated with 1 µg of purified His-AGR2 protein or negative target (His-tag proteins) for pre-clearing step by diluting in 100 µl of coating buffer (0.1 M $NaHCO_3$ pH 8.6) in three wells and incubated overnight at 4^0C. After 4x washes with 200 µl PBST, the wells were then blocked with blocking solution (3% (w/v) BSA in

PBST), rocking for 1 h at RT. Blocking solution was removed followed by addition of 50 µl of the scFv phage library and 50 µl blocking solution to the pre-clearing well containing His-tag protein and incubated 1h at RT with shaking. The scFv library was isolated from a canine naïve phage library provided by Dr Euan Murray. This library encodes both the V_H and V_L chains connected via a linker of 15 amino acids. The phages from the pre-clearing well were transferred into the panning well coated with the target His-AGR2 protein or non-coated well as negative control and incubated for 1 h at RT with shaking. The phages were discarded and wells were washed 10 times with 200 µl PBST. The bound phages were eluted by incubating with 110 µl of 100mM trimethylamine (TEA) for 25 minutes at RT with gentle rocking and neutralised with 22 µl of 1 M Tris pH 7.4. Eluted phages were stored at 4°C for up to 1 week. The neutralised eluted phages were then amplified by infection of *E.coli* TG1 cells. This was done by adding the phage to 1 ml of log-phase *E.coli* TG1 cells and incubated at 37°C for 15 min. The infected cells were transferred into 20 ml LB media supplemented with 100 µg/ml ampicillin and incubated at 37°C for 1.5 h to select for the phage-infected cells. Twenty µl of helper phage M13KO7 (New England Biolabs, N0315S) was added and incubated for 15 min static followed by 1.5 h shaking at 37°C. Kanamycin with a final concentration of 50 µg/ml was added and incubated at 30°C overnight. To precipitate the phages, the overnight bacterial culture was pelleted at 4000 rpm for 10 min. Supernatant containing the phage was added to 3.5 ml polyethylene glycol (PEG) 20% (w/v) in 2.5 M NaCl and precipitated for 1 h on ice. The precipitated phages were pelleted at 4000 RPM for 15 min. The supernatant was discarded, re-centrifuged briefly and any remaining liquid was removed. The pellet was resuspended in 200 µl 1x PBS and used for the next biopanning round.

The biopanning procedure was repeated two times to enrich for scFv binding to target. The precipitated phages were diluted 2:100 in blocking solution for each biopanning round and repeated as described above except after the incubation with the target; the wells were washed 20x instead of 10x with 200 µl PBST.

2.9.2. Isolation of scFv clones and expression of scFv

The neutralised phage pools from each panning round were tittered by making serial dilution to give 10^{-3}, 10^{-6} 10^{-7}, and 10^{-8} dilutions. From each dilution, 2 µl were infected to 98 µl *E.coli* TG1 cells culture and incubated for 15 min at 37^0C static. Fifty µl of the infected cells were plated into LB agar plates with ampicillin 100 µl/ml and incubated overnight at 37^0C incubator. Colonies were picked and grown into 100µl LB media with ampicillin 100 µl/ml (LB-amp) into a 96 well plate (Corning tissue culture plate) and grown for 4 hours at 37^0C. The culture was diluted 1:100 in LB-amp by adding 8µl culture into 792µl in a 96-deep well plate. To the remaining culture in the plate, sterile 80% glycerol was added to the first plate, and this plate was kept at -80^0C to reuse in the future (scFV glycerol stocks). The culture was grown for two hours at 37 ^{0}C before adding 200µl of LB-ampicillin with 0.1% glucose (w/v) containing 5 mM IPTG (final concentration 1 mM) per well and grown overnight at 30ºC. The plates were spun down at 4000rpm for 10 minutes to obtain a supernatant containing secreted scFv.

2.9.3. Detection of scFV phage pools or scFv clones binding by ELISA

A 96-well polystyrene assay plate (Corning™ Costar™ 96-Well white solid, #07-200-592) was coated with 0.5 µg of purified His-AGR2 protein or non-coated (coating buffer only) in 50 µl of coating buffer (0.1 M $NaHCO_3$ pH 8.6) in triplicates and incubated overnight at 4^0C. The wells were washed 4x with 200 µl PBST and incubated in blocking solution (3% (w/v) BSA in PBST), rocking for 1 h at RT. The precipitated phage-scFv pool or individual phage-scFv clone was diluted 1:50 in 50 µl blocking solution per well and incubated 1 h at RT shaking. The wells then were washed 6x with 200 µl of 0.1% PBST followed by the addition of anti-M13-HRP monoclonal (GE Healthcare, #27942101) conjugated antibody at 1:5000 dilution in blocking solution for 1 hour RT shaking. The wells of the plate were washed again 6x with 200 µl of 0.1% PBST to remove excess primary antibody and finally detected by adding ECL and reading using Fluoroskan Ascent FL (Labsystems) as relative light unit (RLU).

2.9.4. Extraction of phagemid for sequencing

Glycerol stock of scFv clones stored at -80^{0}C was inoculated into 5 ml LB-amp and grown overnight (16-18 hours) at 37^{0}C shaking. Alternatively, 1 µl of precipitated phage was infected into 1 ml log-phase *E.coli* TG1 cells and incubated in 37^{0}C static and incubated 15 minutes. Five ml of LB-amp were then added to the infected cells and incubated overnight at 37^{0}C overnight (16-18 hours). The phagemid was extracted using QIAprep Spin Miniprep Kit (Qiagen) as described in 2.3.1. The phagemid was sequencing using a gIIIFor reverse primer as described 2.3.5 to obtain a sequence of scFv.

2.10. Gene editing using CRISPR/Cas9 technology

2.10.1. Guide RNA design

Guide RNAs (gRNA218 and gRNA219) for AGR2 knockout were obtained from Horizon Discovery and were cloned into a GE6-002 plasmid. This plasmid contains a U6 promoter, expressing wild-type Cas9 which is under the control of CMV promoter and also contains a 2A puromycin resistance gene for selection. The gRNA designed for AGR2 knockout was also designed for AGR2 knockin mutation (AGR2 C81S), where this mutation will have defective oxidoreductase activity on its sole cysteine within the thioredoxin domain. The AGR2 knockout/knockin gRNA targets exon 4 of AGR2 (as the DNA sequence containing the Cys81 is in exon 4).

2.10.2. Transfection of gRNA, expansion, and characterisation of clones

MCF-7 cells were transfected with 1.2 µg of gRNA construct using Attractene as described in 2.4.3 a 6-well plate. The transfected cells were incubated 72 hours. The transfected cells were serially diluted and plated into 96-well plates at 1 cell per well in standard cell culture media. Ten 96-well plates containing single cell dilution were obtained. After several days, each well of 96-well plates were checked and marked if they contained a single colony, while wells containing more than one colony (which may contain contaminating clonal line) were excluded. The clonal lines were then incubated in the incubator for several weeks to expand and the media were changed every 3-5 days. The clones were expanded to 24-wells, then to 12-wells, and finally to

6-wells plates. Expanded clonal lines were harvested and analysed using western blot or genotyped using SURVEYOR nuclease assay described below.

2.10.3. Functional validation of clonal line

SURVEYOR® mutation detection kit (Integrated DNA technologies # 706025) was used for genotyping AGR2 knockout clonal line or a pool of transfected cells with gRNA. First genomic DNA was isolated using Gentra Puregene Cell Kit (Qiagen #158745) following the manufacturer's directions. SURVEYOR PCR primers for AGR2 knockout were designed from the genomic DNA region within AGR2 exon 4 that will amplify 586 bp long, where one of the primers is roughly one-third of the Cas9 target so that the cleavage fragments can be clearly distinguished in electrophoresis (Table 2.26). PCR was performed as described in 2.3.3.1 using the condition in Table 2.27. PCR products were cleaned up by QIAquick PCR Purification Kit before sending for sequencing using AGR2 Exon 4 F as a primer (Table 2.26) or genotyping using SURVEYOR® mutation detection kit as per manufacturer's protocol. Digested PCR fragments from SURVEYOR® assay were analysed using 2% DNA agarose gel electrophoresis.

Plasmid	**Sequence 5'-3'**
AGR2 Exon 4 F	TCTTTAGTATAGGGTCTATTATCCCAGTCT
AGR2 Exon 4 R	TATTGAATGTGGACTGTAAGAGAAATCAAG

Table 2.26 SURVEYOR primers for the AGR2 knockout.

PCR cycles protocol		
Step 1	95°C	2 min
Step 2	95°C	20 sec
Step 3	60°C	30 sec
Step 4	72°C	1 min
Step 5	Repeat step 2-4 30 times	
Step 6	72°C	5 min
Step 7	4°C	∞

Table 2.27 PCR cycles protocol for SURVEYOR nuclease assay.

2.11. Hydrogen-deuterium exchange mass spectrometry

2.11.1. Proteins used for hydrogen-deuterium exchange mass spectrometry.

The amino acid sequences of the AGR2 recombinant proteins with N-terminal His-tag are:

1) wild-type AGR2:
MSYYHHHHHHLESTSLYKKAGFEGDRTMRDTTVKPGAKKDTKDSRPKLPQ
TLSRGWGDQLIWTQTYEEALYKSKTSSKPLMIIHHLDECPHSQALKKVFAEN
KEIQKLAEQFVLLNLVYETTDKHLSPDGQYVPRIMFVDPSLTVRADIT

2) mutant AGR2^{S134A}
MSYYHHHHHHLESTSLYKKAGFEGDRTMRDTTVKPGAKKDTKDSRPKLPQ
TLSRGWGDQLIWTQTYEEALYKSKTSSKPLMIIHHLDECPHSQALKKVFAEN
KEIQKLAEQFVLLNLVYETTDKHLSPDGQYVPRIMFVDPALTVRADITGRYS
NRLYAYEPADTALLLDNMKKALKLLKTEL.

The proteins were purified as described above and gel-filtered into a buffer with the composition of 20mM Tris pH8.0, 150mM NaCl and 10mM $MgCl_2$. The 16mer AGR2-interacting peptide has the amino acid sequence SGSG-HLPTTIYYGPPG and the stock solution is at 5 mg/ml in DMSO. The peptide: AGR2 protein ratio was at 10:1 to give a final concentration of the AGR2-binding peptide at 10 μM and a final concentration of DMSO was 1%.

2.11.2. Sample preparation and statistical rationale for hydrogen-deuterium exchange mass spectrometry

Deuteration of the His-AGR2 WT and His-AGR2^{S134A} proteins was initiated by a sequential 10-fold dilution into a deuterated buffer (20 mM Tris-HCl, pD 7.6, 150 mM NaCl, and 10 mM $MgCl_2$). In the peptide mapping experiment, either His-AGR2 WT at 3 μM final concentration was incubated with a final concentration of peptide at 10 μM [181]. This higher concentration of peptide was used as we had previously defined the K_d to be 45 μM using fluorescence polarisation [181]. Deuteration of full-length MDM2 without or with Nutlin-3 was performed as described previously [182]. Samples were processed in biological replicates with

representative data from a time course of deuteration (AGR2 without and with peptides or wt-AGR2 compares to the $AGR2^{S134A}$ mutant) highlighted in each experiment. Multiple peptide replicates were used to selected data and the HD Examiner software selects peptides with high confidence meaning the selected peptides are identified in all submitted MS data of all samples (non-deuterated and deuterated time intervals). When peptides are not identified in all samples, then the deuteration of the peptide is not calculated. Protein was incubated with peptide for 30 minutes prior to the exchange. The deuterium exchange was carried out at room temperature and was quenched by the addition of 1 M HCl in 1 M glycine at 30 s, 1 min, 3 min, 10 min, 30 min, 1 h, and 3 h followed by rapid freezing in liquid nitrogen. Each sample was thawed and injected onto an immobilised pepsin column (15 µl bed volume, flow rate 100 µl/min, 2% acetonitrile / 0.05% trifluoroacetic acid). Peptides were trapped and desalted on-line on a peptide microtrap (Michrom Bioresources, Auburn, CA) for 3 min at flow rate 50 µl/min. Next, the peptides were eluted onto an analytical column (Jupiter C18, 1.0 x 50 mm, 5 µm, 300Å, Phenomenex, CA) and separated using a linear gradient elution of 10 % B for 2 min, followed by 17 min isocratic elution at 40 % B. Solvents were: A – 0.1% formic acid in water, B – 80% acetonitrile / 0.08% formic acid. The immobilised pepsin column, trap cartridge and the analytical column were kept at 1°C. Mass spectrometric analysis was carried out using an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific) with ESI ionisation on-line connected with a robotic system based on the HTS-XT platform (CTC Analytics company). The instrument was operated in a data-dependent mode for peptide mapping (HPLC-MS/MS). Each MS scan was followed by MS/MS scans of the top three most intensive ions from both CID and HCD fragmentation spectra. Tandem mass spectra were searched using SequestHT against the cRap protein database (ftp://ftp.thegpm.org/fasta/cRAP) containing sequences of the AGR2 proteins with the following search settings: mass tolerance for precursor ions of 10 ppm, mass tolerance for fragment ions of 0.6 Da, no enzyme specificity and no fixed or variable modifications were applied. The false discovery rate at peptide identification level was set to 1%. Sequence coverage was analysed with Proteome Discoverer software version 1.4 (Thermo Fisher Scientific). Analysis of deuterated samples was done in HPLC-MS mode with ion detection in the orbital ion trap and the data were processed

using HDX Examiner (Sierra Analytics). Graphs summarising deuteration kinetics were plotted using the Draw H/D Protection Plot [183]. All peptide plots (graphs showing % deuteration over a time course) are summarised in Appendix.

2.12. Tandem mass tag (TMT) Mass Spectrometry

2.12.1. Sample preparation

FLO-1 cells were grown to 90% confluency. FLO-1 cells were split, counted and transfected in biological triplicates by electroporation using a Nucleofector® Device (Lonza) and the Amaxa Cell Line Nucleofector Kit T (Lonza; # VCA-1002) according to manufacturer's instructions. Briefly, 2x 10^6 cells were used with 2 µg of each DNA constructs; 1) pAM1L (FL-AGR2 WT/RSV 2) pAM2L (FL-AGR2 KDEL/RSV) or 3) pSF-RSV vector only. For the electroporation a high expression program (U-023 on Nucleofector® Device) was used. Electroporated cells were plated into a 6-well plate, allow to recover and the media was changed the following day. The cells were further incubated for 72 hours post-electroporation before harvesting. Media were removed, washed 3 times with PBS and harvested using urea lysis buffer (8M urea, 0.1 M Tris-HCl pH 8) by incubating 30 mins on ice. Cellular debris was spun down 14 000 rpm for 20 min at 4^0C. Cell lysates were then subjected to tryptic digestion using filter-aided sample preparation (FASP) digestion protocols as previously described [184]. The filter unit (Hydrosart—stabilised cellulose-based membrane; Vivacon 500 µl (Sartorius Stedim Biotech, #VN01H02 with a 10 kDa cutoff). The filter unit was washed with 100 µl of same urea lysis buffer from harvesting, centrifuged at 14 000 rpm for 15 min at RT. Two hundred µg of the protein sample in urea lysis buffer was added to the filter unit followed by centrifugation for at 14 000 rpm for 15 min at RT. Urea lysis buffer containing 16.7 mM tris 2-carboxyethly phosphine hydrochloride (TCEP) was then added followed by incubation in a thermomixer for 600 rpm for 30min at 37˚C followed by centrifugation at 14 000 rpm for 15 min at RT. This is to reduce any disulphide bonds in the denatured proteins. To alkalize the proteins of free sulfhydryl groups to form S-carboxyamidomethyl-cysteine for maximal tryptic digestion, the filter unit was incubated with urea lysis buffer containing 50 mM iodoacetamide in the dark for 20 min at room temperature before further centrifugation at 14 000 rpm for 15 min. The filter unit was washed with

a buffer containing 100 mM ammonium bicarbonate (NH_4HCO_3), centrifuged at 14 000 rpm for 15 min at RT A buffer containing 50 mM NH_4HCO_3 (100 µl) and trypsin (Promega) in a mass ratio 1:30 (trypsin/protein) was added to the filter, the samples were mixed at 600 rpm for 1 min, and then incubated for 18 h in a wet chamber (37˚C). Two rounds of subsequent centrifugation 14 000 rpm/15 min (room temperature) eluted the peptides. Samples were centrifuged at 14000 g at 20^0C for 15 minutes, the filters were discarded. The concentration of the digested peptides was measured using the Quantitative Colorimetric Peptide Assay (Pierce, ThermoFisher) following manufacturer's instruction. The absorbance of the reaction was measured using a TECAM Spectrophotometer at 480 nm. A calibration curve was created and peptide concentration was determined. 25 µg of peptides were transferred in a new Eppendorf and dried in a SpeedVac at 35^0C for 1 hour or until the liquid had evaporated completely. Samples were stored at -80^0C.

2.12.2. TMT labelling

For the labelling of the peptides, the Thermo Scientific Tandem Mass Tag™ Labelling kit (TMTsixplex™ Isobaric Label Reagent Set # 90061) and reagents were used. This labelling allows the multiplex relative quantitation by mass spectrometry. The reagent set was used to label 6 different samples as described in Table 2.28. For each of them, a unique reporter mass in the low mass region of the MS/MS spectrum is used to measure relative protein expression levels during peptides fragmentation.

Label Reagents	Sample ID
TMT^6 -126	RSV CONTROL 1 (pSF_RSV)
TMT^6 -127	RSV CONTROL 2 (pSF_RSV)
TMT^6 -128	AGR2 WT 1 (pAM1L)
TMT^6 -129	AGR2 WT 2 (pAM1L)
TMT^6 -130	AGR2 KDEL 1 (pAM2L)
TMT^6 -131	AGR2 KDEL 2 (pAM2L)

Table 2.28 TMTsixplex isobaric label reagent set.

Twenty-five µg of peptides per sample were labelled with TMT labels, according to Thermo Fisher Scientific protocol. Peptides were separated using Thermo Scientific

UltiMate 3000 RSLCnano chromatograph. Sample was loaded on a precolumn (Thermo Scientific µ-precolumn, 30 µm id, 5 mm length, C18 PepMap 100, 5 µm particle size, 100 Å pore size) and peptides further separated on a Thermo Scientific™ Acclaim PepMap RSLC column (75 µm id, length 500 mm, C18, particle size 2 µm, pore size 100 Å) using a 300 nL/min flow rate with a linear gradient of B (80% ACN in 0.08% aq. formic acid) in A (0.1% aq. formic acid (FA)). The gradient composition used for peptides separation was as follows highlighting time as a function of % B: 0 min, 2%B; 10 min, 2%B, 200 min, 50%B, 210 min, 98%B, 228 min, 98%B, 255 min END. peptides were transferred into a newly labelled microcentrifuge tube and evaporated using a SpeedVac concentrator and were stored at −80◦C. Dried, desalted peptides were reconstituted in 0.1% FA and introduced into Thermo Scientific™ Orbitrap Elite™ operating in Top10 data-dependent acquisition mode. The data acquisition parameters setting for the Top10 method include: MS1 includes a mass range of 335–1800, a resolving power setting of 120 000, and a Max. in a time of 200 ms; MS2 includes: HCD activation type; a min. The signal required of 5000, an isolation width of 1.2, NCE of 35, charge rejection state is unassigned; 1+; dynamic exclusion settings are ON; the exclusion width is 10 ppm; repeat count is set at 1; the exclusion list size is 500, and the exclusion duration is 30 s.

2.12.3. Database searching and analysis

Data analyses were processed with a Proteome Discoverer1.4 (Thermo Scientific™), employing Mascot with the following search settings: database Swiss-Prot human (April 2015); enzyme trypsin; two missed cleavage sites; precursor mass tolerance 10 ppm; fragment mass tolerance 0.6 Da; dynamic modifications: formyl [peptide N-terminus], oxidation [M], Gln to pyro-Glu [peptide N-terminus], acetyl [protein N-terminus], TMT6plex [K], TMT6plex [N-term], and static modification: carbamidomethyl [C]. The results of the search were further submitted to generate the final report using a cutoff of 1% FDR on peptide levels, and only unique peptides were used for protein quantitation. The quantitative option was enabled with the corresponding combination of labelled peptides in the sample type and the observed relative quantification ratio was normalised compared to the median.

CHAPTER 3: Defining a Function for the Sequence-Specific Peptide Docking Site by the Pro-Oncogenic Protein AGR2

3.1. Introduction

AGR2 protein has emerged in the past ten years as a major pro-oncogenic and metastatic player in cancer largely due to the emergence of clinical-OMICS [94] [116]. AGR2 protein is highly upregulated in a various range of human cancers and has been shown to be involved in asthma, inflammatory bowel disease, cell transformation, cancer drug resistance and metastatic growth as previously discussed. However, the biological function of AGR2 is not fully understood, and scientists were unable to embed AGR2 in a specific pathway. Finding evidence of multifaceted roles of AGR2, and understanding its biological functions could cast mechanisms by which AGR2 contributes to diseases. One way to understand the diverse biological functions of AGR2 is to understand how AGR2 interact with its client proteins. By suggesting features of binding partners, they provide a cellular framework that can give mechanistic insights about function. Current goals of this emerging AGR2 field are, therefore, to characterise and identify protein interactions that mediate its function which include identification of the signalling pathways through which it can stimulate cell growth or drug-resistance[185, 186], and determining whether these pathways are "druggable"[187].

3.1.1. Linear peptide motif as tool to identify AGR2 binding protein

There are approximately 22,000 protein-coding genes in the human genome which translated to roughly 35,000 instances of folded domains belonging to about 10,000 families [188]. Their functions are often grasped through interactions with other proteins, mediated by an estimated 10,000 distinct interaction types [189], resulting in several hundred thousands of binary interactions [190-192] and 600–1,000 distinct protein complexes [193-195]. Proteins have been assumed classically to bind each other via large discrete region called domain. However, with the recent advances in bioinformatics and genome sequencing have shown that majority of protein-protein

interactions in nearly all cellular processes are mediated by small interaction modules, referred to as peptide motifs (Figure 3.1). Peptide motifs are typically less than ten residues in length signify functional sites that often occur within intrinsically disordered regions of proteins, and there might be a million instances of peptide motifs in human proteome [196].

Peptide motifs are also referred to as linear motifs [LMs] or short/ eukaryotic linear motifs [ELMs/SLiMs]) in the literatures. Peptide motifs were originally discovered as conserved regions in rapidly growing regions and a substantial number of motifs with important functions are protected against deleterious mutations [197, 198]. Peptide motifs can be categorised into two different classes which are binding motifs and post-translational modification sites (PTM). Regulatory role of the binding motif is to mediate interactions with globular domains which often function in the assembly of protein complexes as scaffolding modules. Due to its small size, they are widespread in key biological processes and are estimated to be millions in nature. Binding motif also can be further categorised into functional subsets which include trafficking motifs, targeting motifs, docking motifs and degron motifs. PTM are sites that only recognised when modifying enzymes alter them. PTM sites normally have weaker intrinsic specificity determinants [199, 200] and can be detected in a single residue range. PTM sites are classified into three sub categories which include moiety addition sites, cleavage sites and structural modification sites [196].

Peptide motifs mediated interactions have very dissimilar properties from those interactions between two globular domains. Peptide motifs normally have low affinity for their binding partners and bind to much smaller interface areas [201, 202]. These properties help to clarify why peptide motifs frequently involved in signalling and other regulatory transient interactions and thus give them plasticity in the vast protein-protein networks and lend themselves to rapid cellular perturbation and environmental changes [201]. They can either be switched on or off relatively easily, and their redundancy and weak affinity probably means that the loss or gain-of-function will rarely affect the system.

Although peptide motifs have the weaker disposition, interactions mediated by peptide motifs demonstrate high specificity with their binding partners. This can be achieved by eliminating certain interactions through the absence of other partners and the regions adjacent to peptide motifs. It has been shown that peptide motif flanking regions contribute 21% of the binding energy on average [203]. Further analysis demonstrated that flanking regions are often structurally conserved [204]. These flanking regions help peptide motifs fold into stable structure upon binding to their client domain. Mohan and colleagues described two types of structures adopted by peptide motifs upon binding [205]: formation of α helices, β strands; and formation of irregular structure elements. These structures normally involve the cooperation of multiple interactions and simultaneous presence of multiple weak low-affinity interactions that eventually gives structural rigidity to the overall complex. Peptide motifs can be an important tool in mediating the discovery of protein-protein interaction. Identifying peptide motifs that bind to proteins help to elucidate a group or class of proteins with which they interact, and thus determining their interactomes [206]. The fact that peptide motifs are rarely conserved makes them difficult to find. Most laboratory approaches such as the classical proteomics tandem affinity purification-tag (TAP-tag), yeast two-hybrid (Y2H) or combinatorial peptide-phage display experiments, are typically biased toward stable interactions [202, 206-208].

The AGR2 protein was previously screened for peptide binding aptamers using combinatorial peptide-phage libraries. Briefly, phage-peptide display involves the expression of a peptide of the major coat protein of a phage virion which usually represented by an M15 or filamentous phage) (Phage display procedures is discussed further in details in Chapter 6). Large combinatorial phage peptide display libraries, in which a single random peptide is expressed on the surface of each virion can be built up (roughly 10^9 peptide sequences), and this library can be used to screen for potential binding peptides for a given protein or DNA. Using this approach, purified AGR2 was adsorbed onto a solid phase of an ELISA microtiter plate, a combinatorial peptide phage library was added, and unbound peptides were removed by several washing steps. Peptides bound selectively to AGR2 were released using low-pH and these selection and amplification processes were repeated four times to achieve high-affinity

binding peptides. Two peptides were found to bind AGR2 protein in high affinity; peptide A2 with the sequence YPWHHSWTHTTL and peptide A4 with the sequence HLPTTIYYGPPG [209]. These two peptides bind to two different epitopes on the protein [209]. Peptide A2 binds better to AGR2 when the protein is partially denatured or unfolded. On the other hand, peptide A4 binds better to AGR2 when the protein is native. Only peptide A4 was then further characterised as this peptide showed higher binding affinity to clinical tissue biopsy. Synthetic biotinylated peptide aptamers bound in an ELISA format to AGR2 and substitution mutagenesis further minimised one polypeptide aptamer to a hexapeptide core PTTIYY.

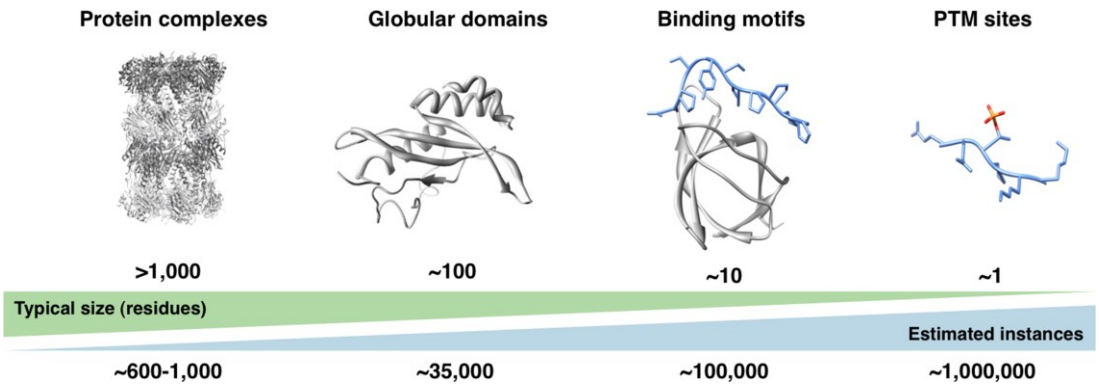


Figure 3.1 Functional Modules of Proteins. (Adapted from [196]) Typical length scale and estimated the number of cases of whole-protein elements (complexes and structural domains [folds]) and protein parts (binding motifs and modification sites) in the human proteome.

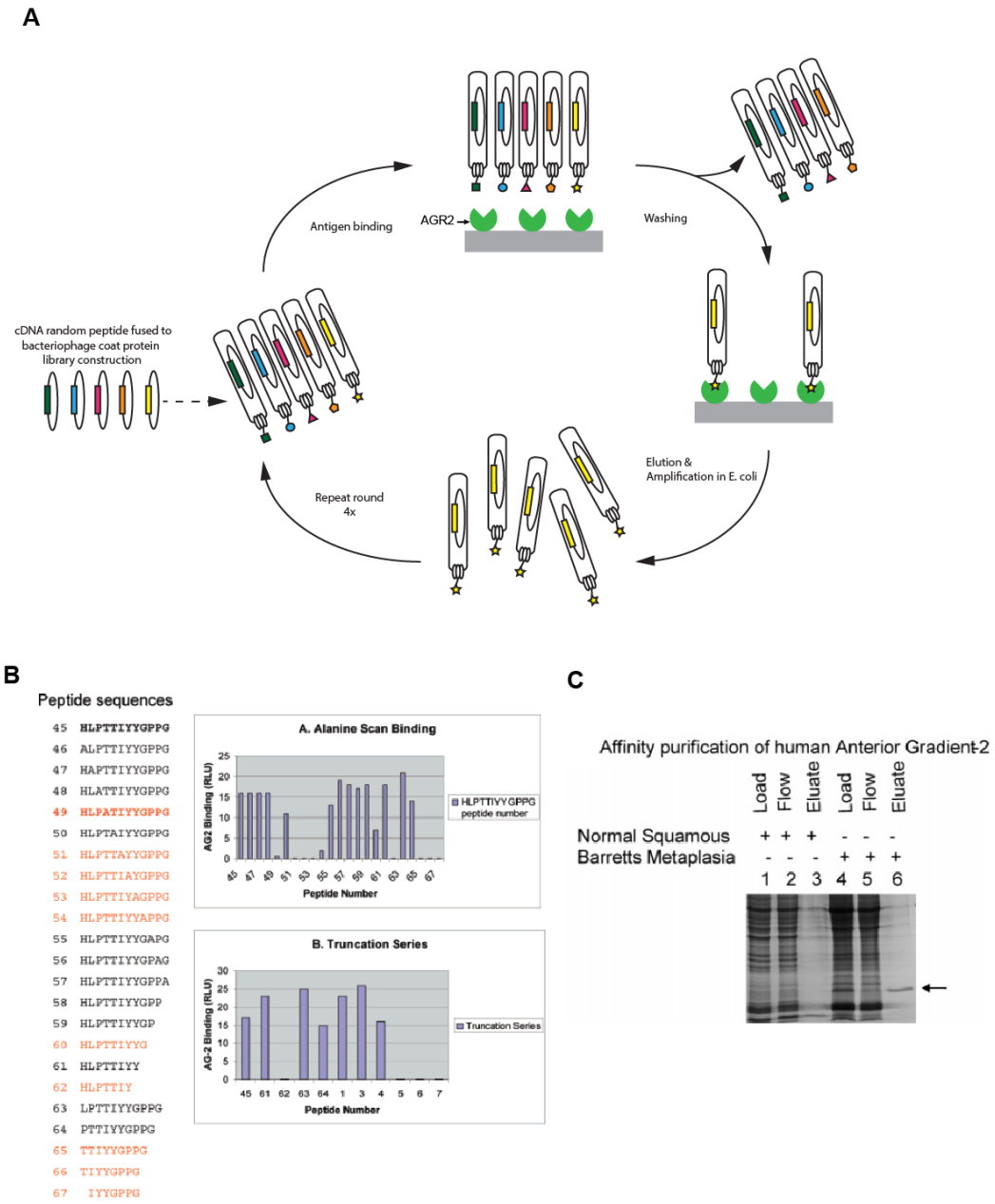


Figure 3.2 Phage-peptide display identifies peptide that binds to AGR2 in-vitro. (A) The phage-peptide library was made by fusing random cDNA peptide to bacteriophage coa protein on the surface of the phage virion. Some peptides bind biotinylated to AGR2 adsorbed into wells of microtiter plate and- non-bound virion peptides are washed away. Bound virions are recovered and used to infect *E.coli* for amplification for use for the next round. Successive rounds of binding, selection, infection and replication enrich for peptides with high-affinity binding. (B) AGR2 binding peptide HLPTTIYYGPPG was identified and subjected to alanine scan and truncation series to find minimal binding peptide [209] (C) Coomasie Blue detection of the minimal binding hexapeptide (PTTIYY) incubated from Barret's esophagus tissue crude lysate (containing AGR2 proteins) which bind AGR2 to near homogeneity .

3.1.2. Aims of chapter

The function of the AGR2 sequence-specific binding peptide has not been defined. In this chapter, we begin to investigate the biology of AGR2 by probing the specific peptide (peptide A4) binding function of AGR2 to define a possible cell biological role for this activity. Hydrogen-deuterium exchange mass spectrometry was employed to determine whether a specific peptide docking site can be mapped on AGR2 followed by generation of 'docking site' mutants. Subsequently, we developed AGR2 consensus peptide-binding motif in which we used it to search for human proteins containing this motif as potential AGR2 client proteins. We then targeted a key protein of interest as putative AGR2 client protein and perform *in-vitro* protein-protein interaction assays to validate AGR2 binding. Accordingly, we proposed a model of how peptide-binding function directs the interaction of AGR2 with its client proteins.

3.2. Results

3.2.1. Mapping AGR2 peptide docking sites using hydrogen- deuterium exchange mass spectrometry.

3.2.1.1. Specificity of peptide A4 binding to AGR2

Peptide A4 with the sequence HLPTTIYYGPPG was previously shown to bind AGR2 with high affinity [209]. We further evaluated the specificity of the peptide to see whether it can bind to and affinity purifies monomeric or dimeric AGR2 protein from human cell lysates. MCF-7 cells which express endogenous AGR2 were left untreated or treated with a cell-membrane permeable cross-linker that stabilises AGR2 dimer via a K95-K95 homodimeric cross-link [97]. When lysates from untreated cells are incubated with streptavidin beads coated with peptide A4, monomeric AGR2 protein can be affinity purified from the crude lysate (Figure 3.3A, lane 4 vs Load, Flow-though, and wash, lanes 1-3). In addition, the chemically stabilised AGR2 dimer (Figure 3.3A, lane 5) can also be affinity purified using the peptide A4 coupled beads (Figure 3.3A, lane 8 vs 5-7).

We then further assessed the specificity of the peptide in affinity capture of AGR2 protein from beads without peptide (Figure 3.3B, lane 2) and beads with peptide (Figure 3.3B, lane 3). These data indicate that peptide A4 can be used to affinity purify dimeric or monomeric protein and suggest the bioactivity of this peptide motif in crude lysates to AGR2 is relatively specific. However, the silver stain gel does reveal high molecular mass proteins that can bind to the peptide directly or through association with AGR2. This fraction containing potential AGR2 binding proteins were not evaluated. Nevertheless, the relative specificity of this peptide for AGR2, as defined by its ability to affinity purify the protein from crude lysates, suggests that this feature of AGR2 can selectively drive (some of) its protein-protein interaction functions. Thus, we continued to characterise this specific peptide binding activity in an attempt discover bonafide client proteins which harbour this peptide motif.

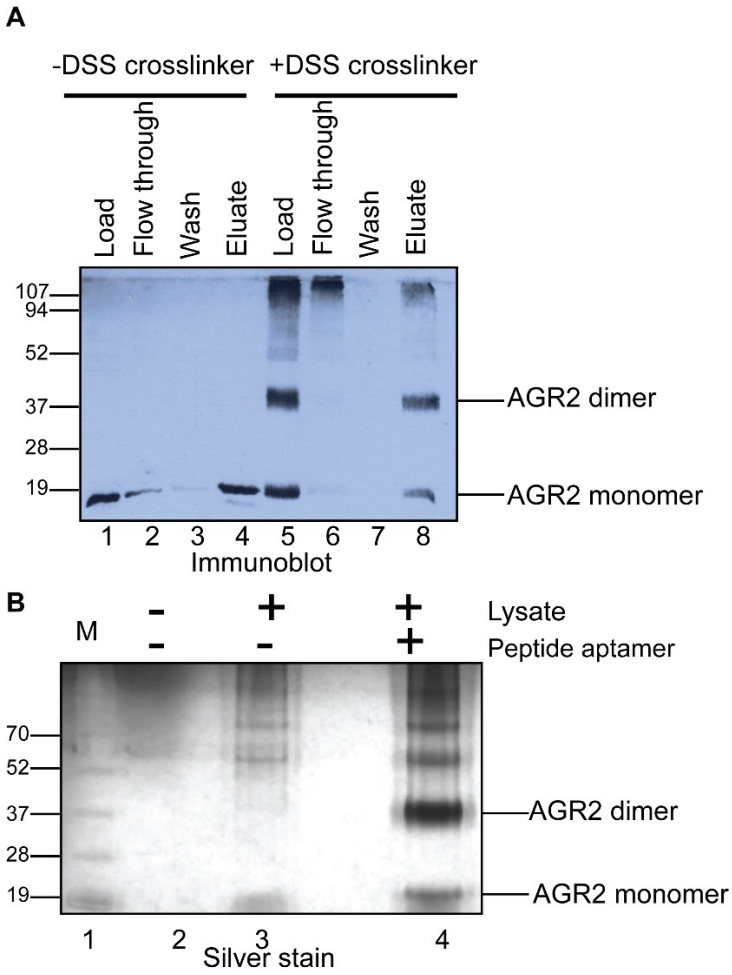


Figure 3.3 Peptide aptamer affinity purification of *in vivo* cross-linked AGR2 protein from crude lysates. MCF7 cells grown in media with 10% FCS were incubated with DMSO (lanes 1-4) or with a fixed concentration of DSS (lanes 5-8) for 1 hour at 37ºC. Cells were harvested, lysed using a Tris-HCl (pH 8.0) buffer containing 1% NP-40, lysates were incubated with an optimised biotinylated peptide aptamer (named A4) linked to streptavidin beads, which can be used to affinity purify the AGR2 protein in crude cell lysates (18). The input lysate (Load, lanes 1 and 5), flow-through (lanes 2 and 6), wash (lanes 3 and 7), and eluate (lanes 4 and 8) was separated by electrophoresis. The protein in the polyacrylamide gel was (A) immunoblotted to determine whether intermediary species of cross-linked AGR2 could be affinity purified and (B) stained with silver to examine total protein capture. The 18 kDa and 36 kDa cross-linked and affinity purified silver-stained proteins (arrows, lane 3 vs lane 2) were excised using trypsin and AGR2 protein was confirmed present by MALDI-TOF mass spectrometry (data not shown).

3.2.1.2. Hydrogen-deuterium exchange mass spectrometry as tool to map peptide A4 binding to AGR2

Hydrogen deuterium exchange (HDX) is a powerful technique which utilises the exchange of hydrogen atoms in the protein with deuterium (D_2O) when protein is diluted into a solution of deuterium (Figure 3.4A). The approach is based on the fact that hydrogen bonding between amino acid side chains and protein backbone amides retard exchange whereas amides that are located in regions that are not stabilised by hydrogen [210]. The degree of protection is indicative of local structure and dynamics. Mass spectrometry is used for determining the mass shifts induced by the exchange of hydrogens by deuteriums with time, as a consequence of changing environments, structure, and complex formation. Thus, this technique is useful for investigating conformational dynamics and interactions in proteins, protein-ligand and protein-protein complexes. When a target protein solution is diluted in D_2O, the protein mass will increase, owing to the exchange between backbone hydrogens and the deuterium in the solution. In a typical HDX experiment, the relative deuterium uptake rates are measured and compared to the protein under different conditions, most commonly with or without ligands.

In the case of the protein-ligand complex is diluted in D_2O, the extent of HDX decreases for regions involved in ligand binding which occurs due to decrease in solvent accessibility in the bonding regions. By comparing changes in deuterium uptake between ligand-free and ligand-bound states of the protein, ligand binding regions can be identified. The global deuteration also can be assessed by measuring the mass of the protein in the presence and absence of ligand which resolves the overall extent of binding under specific experimental conditions. Binding can be probed in a variety of conditions such as altering salt concentrations, changing pH, adding other molecules (e.g., DNA), or mutating a protein [211].

Our lab has previously set up hydrogen-deuterium exchange using Mouse double minute (MDM2) protein as a model system to define effects of ligand binding on the exchange reaction [182]. Purified full-length MDM2 (1 μM) was incubated with DMSO or the ligand Nutlin-3 (a small molecule inhibitor of the MDM2/p53

interaction) for one hour, diluted in a 20-fold excess of D_2O, and processed with pepsin to define changes in deuteration. The data showed expected suppression of deuteration at the N-terminal drug-binding pocket, with internal suppression from amino acids 245-265 that might reflect allosteric effects of the small molecule inhibitor of MDM2 protein (Figure 3.4B). In this study, we aimed to use the same method to monitor deuteration changes of AGR2 protein with and without ligand (peptide A4) to map where Peptide A4 binds to AGR2.

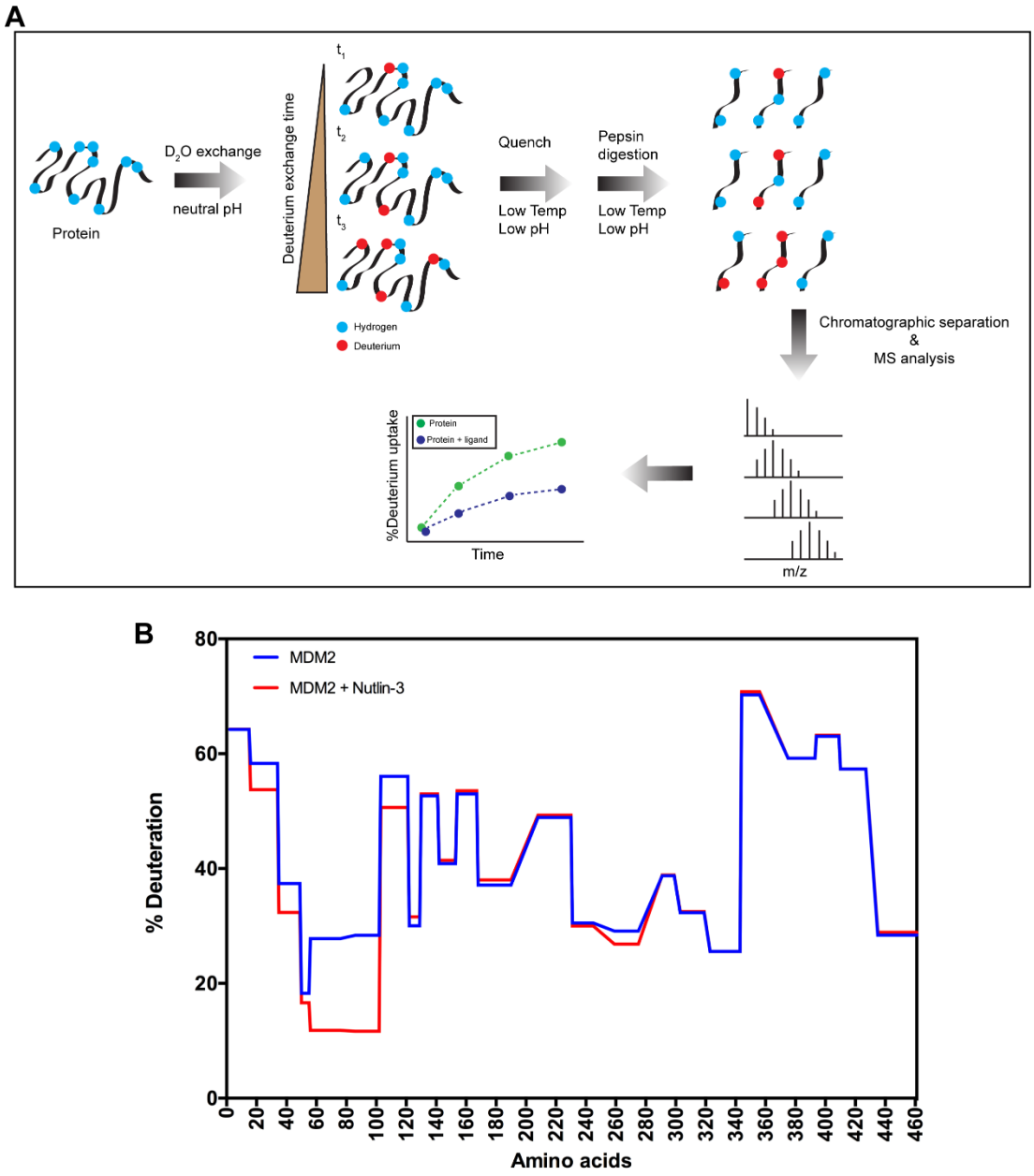


Figure 3.4 Establishing methodology for measuring the effects of a ligand on the target protein MDM2 using hydrogen-deuterium exchange mass spectrometry. MDM2 was assembled with DMSO or with Nutlin-3 in the molar ratio 1 : 4 and cultivated at room temperature for 60 minutes, as described previously for the N-terminal domain of MDM2 (aa 1-126) (26). Deuteration of MDM2 was initiated by sequentially adding D2O with 1% DMSO. The protein concentration during EpCAM/D exchange was 2 µM. The exchange was quenched by acidification to pH 2.3 at eight-time intervals from 40 to 18,000 seconds. The samples were immediately frozen in liquid nitrogen. Thawed samples were processed for pepsinisation as described in the materials and methods. The deuterium exchange rates of individual peptides are summarised using the HDX exchange plots for the 300 second deuteration time course.

3.2.1.3. AGR2 purification for hydrogen-deuterium exchange

In order to perform a hydrogen-deuterium exchange, AGR2 protein needs to be first purified to allow formation complex with its ligand. AGR2 was previously cloned into the pDEST17 vector using Gateway cloning containing hexahistidine tag at N-terminal (Figure 3.5A). As a result, the recombinant protein can be purified using His-tag affinity chromatography. The AGR2 cDNA sequence cloned into the expression plasmid lacked the disordered N-terminal cleavable signal peptide, such that the expressed product would be a representation of the mature form of protein (amino acid 21-175) (Figure 3.5B). Following transformation into *E. coli* BL21 (AI) bacteria and selection, isolated colonies were cultured in LB and induced with IPTG. The induced his-tagged AGR2 fusion product, with a molecular mass of 20 kDa, was evident in crude bacterial lysate (Figure 3.5C, see first lane CL). Following clearing of the bacterial lysate of membrane fragments and unlysed cells, the bacterial slurry was incubated with nickel-agarose beads. Beads were washed and eluted as described. The eluted protein was pooled and buffer exchange to remove imidazole and high salt content. Purified AGR2 was assessed by SDS-PAGE and immunoblot analyses using specific AGR2 polyclonal antibody. By coomassie staining, AGR2 showed a major band at 20 kDa and a minor band around 40 kDa (Figure 3.5C). The 40-kDa band was attributed to a dimeric form of His-AGR2 that persisted even under reductive SDS-PAGE conditions. Overall, following the purification protocol of AGR2, AGR2 was highly expressed with an expected band in bacteria and that the recombinant protein was functional having the N-terminal fusion tag and extra amino acids from the homologous recombination when cloned into pDEST17 using Gateway cloning. The purified AGR2 also is immunoreactive as it can be detected by the AGR2 polyclonal antibody (data not shown).

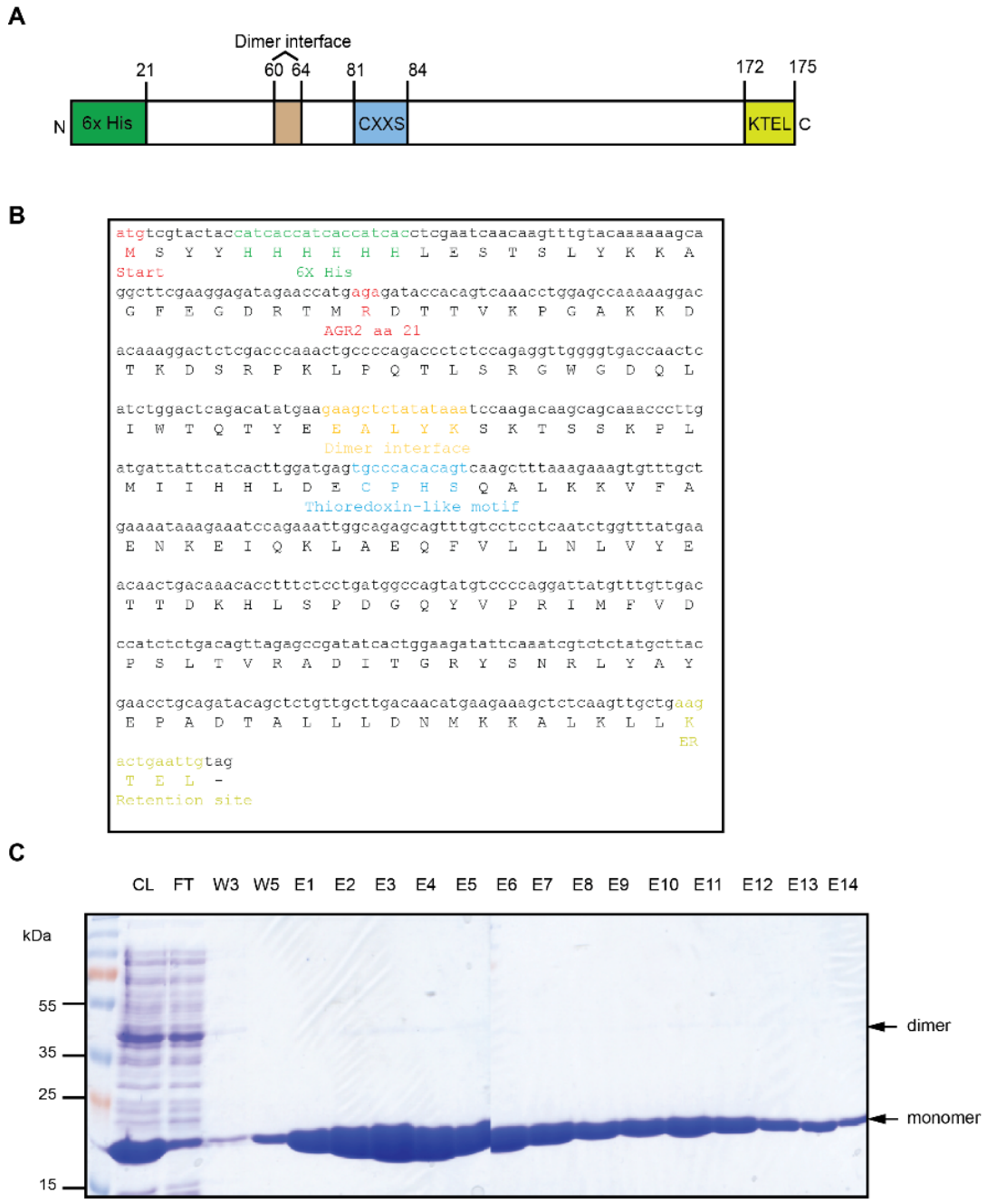


Figure 3.5 Purification of recombinant His-AGR2. A) Schematic representation of His-AGR2 protein and plasmid constructs highlighting essential motifs of N- terminal His-epitope tagged, the dimer interface, thioredoxin motif CXXS and ER retention site KTEL. AGR2 construct for bacterial expression was previously cloned into pDEST17 via Gateway cloning.(B) His-AGR2 DNA and Protein sequences highlighted with AGR2 essential motifs C) Coomassie Blue staining of purified His-AGR2 showing the major band at 20 kDa under reducing SDS-PAGE condition. Diffused AGR2 dimer band can also be seen (arrow).

3.2.1.4. Analysis of deuterium exchange on AGR2 in the presence and absence of peptide-A4

AGR2 protein (3 μM) was incubated without ligand or with peptide A4 (SGSGHLPTTIYYGPPG; 10 μM) to allow complex formation. The samples were then diluted 10 fold in D_2O over a time course from 30 to 10800 seconds, reactions were quenched by acidification and freezing, then the thawed samples were subjected to pepsin digestion (Appendix 1 contains all of the primary raw graphical data). The peptic coverage of a representative experiment using ligand-free AGR2 is shown in Figure 3.6. Peptic cleavage is not completely random throughout the sequence and can generally be grouped into four distinct domains. These domains contain (i) the N-terminal domain from amino acids 21-53 that harbour intrinsically disordered sequences that negatively regulate dimer formation [61] [97]; (ii) the central domain containing amino acids 54-108 that harbor the dimer interface [61], the thioredoxin fold, and the Reptin docking site [105]; (iii) a region N-terminal to the specific peptide-binding domain (aa 109-130; this study); and (iv) the specific peptide binding domain and ER retention site (Figure 3.6).

Preferred cleavage sites or hotspots of pepsin cleavage can be visualized such as Thr23 (highlighted as an N-terminal disordered region; Figure 3.6), Tyr63 (within the dimer interface of 60-EALYK-64), Ile74 (N-terminal to the CxxS motif), and Val110 (highlighted with Reptin binding motif [105]; Figure 3.6). The existence of four regions with a degree of peptic resistance, suggests some secondary/tertiary conformation is maintained upon acidification. It is interesting to note that very few peptides were obtained covering the complete dimerization motif from amino acids 60-64 (Fig 3C; highlighted by dimer interface). In fact, a hotspot of cleavage is at Tyr63 (within the dimer interface of 60-EALYK-64) suggesting this region is relatively more susceptible to proteolysis when in the acidified and denatured state.

Upon peptide binding by AGR2, peptic fragments within the intrinsically disordered N-terminal domain exhibited marginal changes in deuteration (Figure 3.7A-D),suggesting that this region is not significantly impacted upon peptide binding.

Figure 3.6 Peptic peptides recovered from ligand-free mature AGR2 protein (with the N-terminal 20 amino acids containing the hydrophobic ER leader sequence removed. The proteolytic peptide fragments can be grouped into four regions based on extents of recovery after mass spectrometry; (i) The N-terminal region containing the poly his-tag and amino acids Arg21-Leu52 containing the intrinsically disordered region that plays a negative regulatory role in dimer stability (17) (7); (ii) a central grouping from Ile53 to Asn108, containing the beginning of the dimerization motif (60-EALYK-64), the CxxS thioredoxin fold, and the Reptin binding motif (104-FVLLNLVY-111); (iii) a third pepsin resistant region from aa 109-130; and (iv) a region of a cluster of overlapping peptic fragments including the specific peptide binding domain (this study) and the degenerate KTEL endoplasmic reticulum retention site.

Such small mass peptic fragments included the N-terminal tag (YKKAGFEGDRT; Figure 3.7A) and a peptic fragment N-terminal to the dimerization interface (45-SRGWGDQL-52; Figure 3.7B). Interestingly, over the time course of deuteration, the peptide 53-IWTQTYE-59 just adjacent to the dimerization motif was attenuated in its deuteration (Figure 3.7E) and small deuterium suppression over the time course from aa 62-LYKSKTSSKPLM-73 (Figure 3.7F). These latter data suggest that peptide A4 binding can impact on this dimeric interface. There is a degree of specificity in this deuterium suppression around the dimeric interface since the peptide 74-IIHHLDECPHSA-85 containing the thioredoxin fold does not exhibit significant deuteration changes in the presence of peptide A4 (Figure 3.7C). Nor does a small mass peptic fragment adjacent to the Reptin docking site, 97-IQKLAEQF-104, exhibit significant changes in deuteration (Figure 3.7D).

The C-terminal domain showed the most significant global changes in deuteration after peptide binding across the time course (Figure 3.8). For example, peptide 138-RADITGRYSNRL-149 exhibited consistent suppression of deuteration throughout the entire time course (Figure 3.8A) similar to the peptides surrounding the dimer interface. However, the most significantly suppressed minimal peptic fragment over the entire time course of deuteration was 131-VDPSLTVRA-139 (Figure 3.8B). By comparison, the minimal peptide fragment 150-YAYEPADTAL-159 exhibited minimal deuteration changes upon peptide binding (Figure 3.8C). A visualisation of the global changes in peptide deuteration (from Figure 2) deuteration after 30 seconds (Figure 3.8D) or 10,800 seconds (Figure 3.8E) further highlights the most dominant impact of peptide A4 on the structural loop from aa 131-135 (Figure 3.8D-E and A). Presumably, once AGR2-peptide complexes reach some equilibrium after diluting 10 fold with D_2O for 10,800 seconds, there is enhanced deuteration within the dimerization motif (compare Figure 3.8E to Figure 3.8D). This suggests that specific peptide binding might destabilise the dimer into the monomeric state.

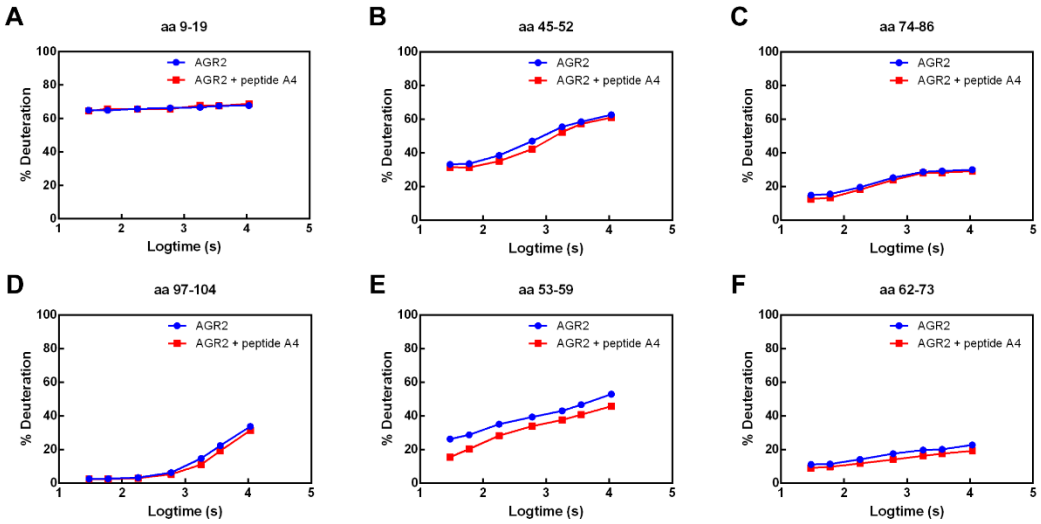


Figure 3.7 Representative deuteration rates of peptides derived from the N-terminal domain of AGR2 after binding of peptide A4. Ligand-free AGR2 or AGR2-peptide A4 complexes were deuterated over a 7-point time course from 30 seconds to 10,800 seconds followed by acidification, pepsinisation, and separation of fragments using mass spectrometry. (A-C). Representative segments of AGR2 protein that do not exhibit significant changes in deuteration after peptide A4 binding over the time course are highlighted including; (A) YKKAGFEGDRT from the N-terminal HIS tag; (B) a peptic fragment N-terminal to the dimerization interface containing aa 45-SRGWGDQL-52; (C) a peptide containing the thioredoxin fold, from aa 74-IIHHLDECPHSA-86; and (D) a peptic fragment adjacent to the Reptin docking site, containing aa 97-IQKLAEQF-104. Segments of AGR2 protein that do show a degree of deuterium suppression over the entire time course included; (E) a peptide containing aa 53-IWTQTYE-59 just adjacent to and N-terminal to the dimerization motif; and (F) peptide containing half of the dimerization motif from aa 62-LYKSKTSSKPLM-73. The data are plotted as % of deuteron exchange as a function of time (log10 in seconds; from 30, 60, 180, 600, 1800, 3600, and 10800).

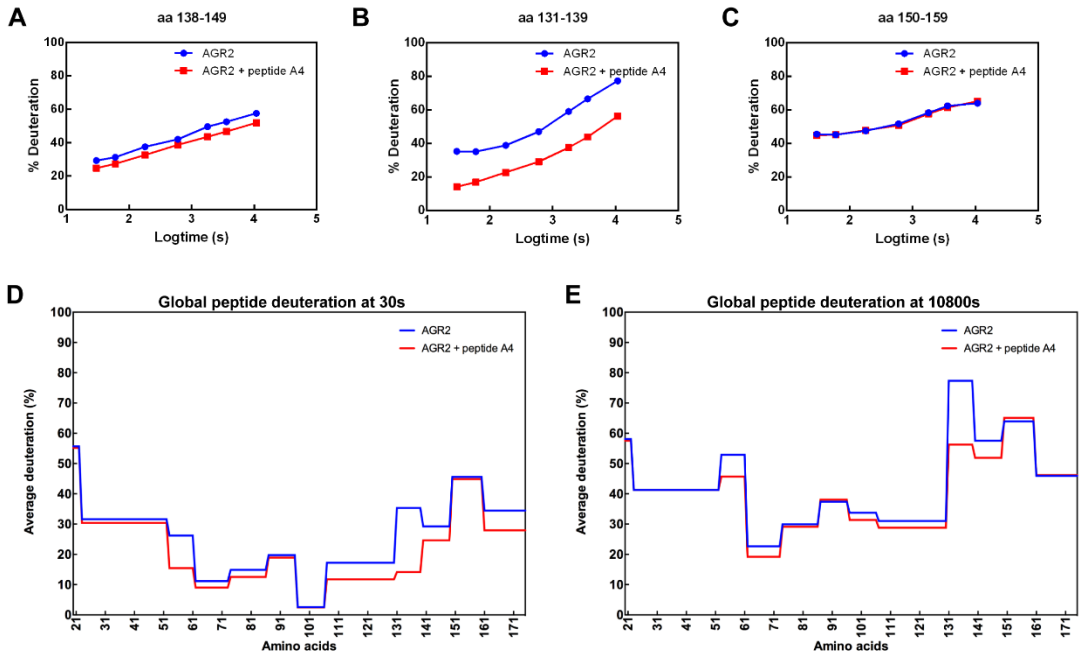


Figure 3.8 Representative deuteration rates of peptides derived from the C-terminal domain of AGR2 after binding of peptide A4. Ligand-free AGR2 or AGR2-peptide A4 complexes were deuterated over a 7-point time course from 30 seconds to 10,800 seconds followed by acidification, pepsinisation, and separation of fragments using mass spectrometry. The C-terminal domain showed the most significant global changes (suppression) in deuteration of individual peptic fragments after peptide A4 binding across the time course are highlighted, including: (A) peptide containing aa 138-RADITGRYSNRL-149, and (B) the peptide contiguous and N-terminal to (A) containing aa 131-VDPSLTVRA-139. (C) The peptide contiguous and C-terminal to (A) containing aa 150-YAYEPADTAL-159 exhibited minimal deuteration changes upon peptide A4 binding. (D and E). A visualisation of the global changes in peptide deuteration without (blue) and with peptide A4 ligand (red) after (D) 30 seconds of incubation in D2O or (E) after 10,800 seconds of incubation in D2O. The data are plotted as % of deuteron exchange as a function of time (log10 in seconds; from 30, 60, 180, 600, 1800, 3600, and 10800).

3.2.1.5. Effect of mutating AGR2 peptide 'docking site'

The most significantly suppressed deuteration was shown in the peptide fragment 131-VDPSLTVRA-139 and this region also showed consistent suppression at early and late time points which most likely to be the binding site for AGR2 peptide motif (Figure 3.9A). Based on the AGR2 crystal structure, the position 131-VDPSLTVRA-139 consists of the two B-sheets and the core VDPSL makes the turn between the B-sheets B). We then decided to mutate the core VDPSL motif (Figure 3.9B) from the his-tagged wt-AGR2 at the positions of D132, P133, and S134 to determine if alterations in PTTIYY peptide binding occur (Figure 3.9C-E). The proteins were purified (Figure 3.9C-E, lower panel) the same way as the wt-AGR2 and tested for changes in consensus site peptide binding. Protein purifications showed the different stability of the mutant's protein in bacteria. $AGR2^{P133A}$ and $AGR2^{S134A}$ were stable and highly expressed and the yield was relatively the same as the wt-AGR2 (Figure 3.9D-E). The mutant $AGR2^{D132A}$, on the other hand, exhibited a high degree of proteolysis and instability in bacteria and its yield was lower than wt-AGR2 suggesting it might have a different conformation (Figure 3.9C, compare the yield of elution fractions). The mutant proteins were purified in at least three different independent experiments and showed relatively same constancy and yield. Despite low yield of $AGR2^{D132A}$, it was still used to evaluate in consensus site peptide binding assays.

Biotinylated peptide A4 was captured on the streptavidin solid phase (Figure 3.9F). The wild-type and docking-site mutated AGR2 proteins were titrated to determine whether these mutations increase, decrease, or have no effect on peptide binding. $AGR2^{S134A}$ exhibited wild-type levels of peptide binding, $AGR2^{D132A}$ exhibited attenuated levels of peptide-binding, and the $AGR2^{P133A}$ exhibited the lowest amounts of peptide binding (Figure 3.9G). The attenuated activities of the AGR2 mutants encoded by the P133A and D132A alleles suggest that these two amino acids play more dominant roles than the S134 side-chain in sequence-specific peptide docking.

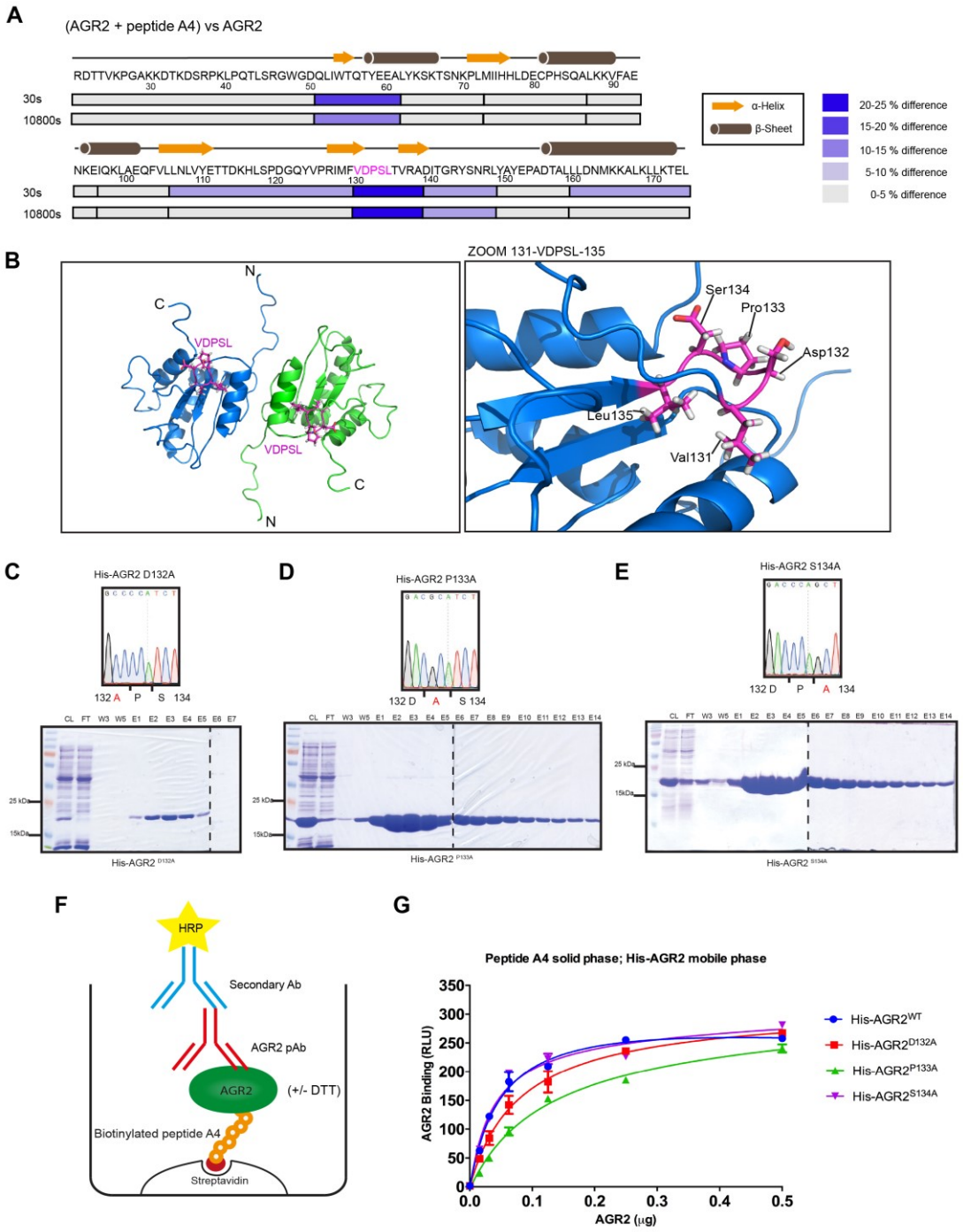


Figure 3.9 Mutations in the dominant deuteration responsive motif impact on AGR2 peptide binding activity *in-vitro*. (A) A secondary structure summary of changes in deuteration in the presence of the consensus peptide. The diagram shows the full amino acid sequence of mature AGR2 protein (aa 21-175) with the alpha helices and β-sheets highlighted. The block colours highlight the changes in fold deuteration after 30 seconds or 10800 seconds, with dominant changes at the region 131-139 (in between two β-sheets) and the dimer interface (aa 50-60). (B) A representation of the minimal deuteration responsive motif from aa 131-135 motif (highlighted in pink) in AGR2 (PDB code: 2LNS)

that is most significantly suppressed by peptide A4 binding. The main amino acids of focus were D132, P133, and S134 flanked by the hydrophobic amino acids V131 and L135. (C-E) DNA sequencing chromatogram traces of three of AGR2 peptide docking site mutations at position Asp132 to Ala, Pro133 to Ala, and Ser134 to Ala (upper panel) and their corresponding protein purifications shown as Coomassie Blue gel showing the relative purity of the indicated mutant proteins expressed in *E.coli* after nickel affinity purification (lower panel). (F-G) An ELISA assay was developed to measure the binding of AGR2 to synthetic biotinylated peptide A4 captured on the streptavidin coated solid phase. Reactions were added to the solid phase to measure binding to biotinylated peptide A4 on the solid phase. The data plot the binding of AGR2 in relative light units (RLU) as a function of increasing wt or mutant AGR2 protein isoforms, as indicate (in µg) that was quantified using AGR2 specific antibody. 12-mer negative peptide F3 which showed no binding specificity to AGR2 from previous studies was used as negative control and ELISA (data not shown). The experiment was carried out in triplicate in three independent experiment, and data is representative of twice replicated experiments.

The results may suggest that aspartic acid (D) residue which is negatively charged amino acid is essential for AGR2 binding and substituting it with uncharged alanine, significantly perturb the binding of AGR2. Proline (P) residue is a unique amino acid that can function to introduce a kink in alpha helices. Proline substitution to alanine is thought to remove the 'kink effect' or 'binding pocket' existed in the wt-AGR2 thus showed loss-of-function binding to the peptide.

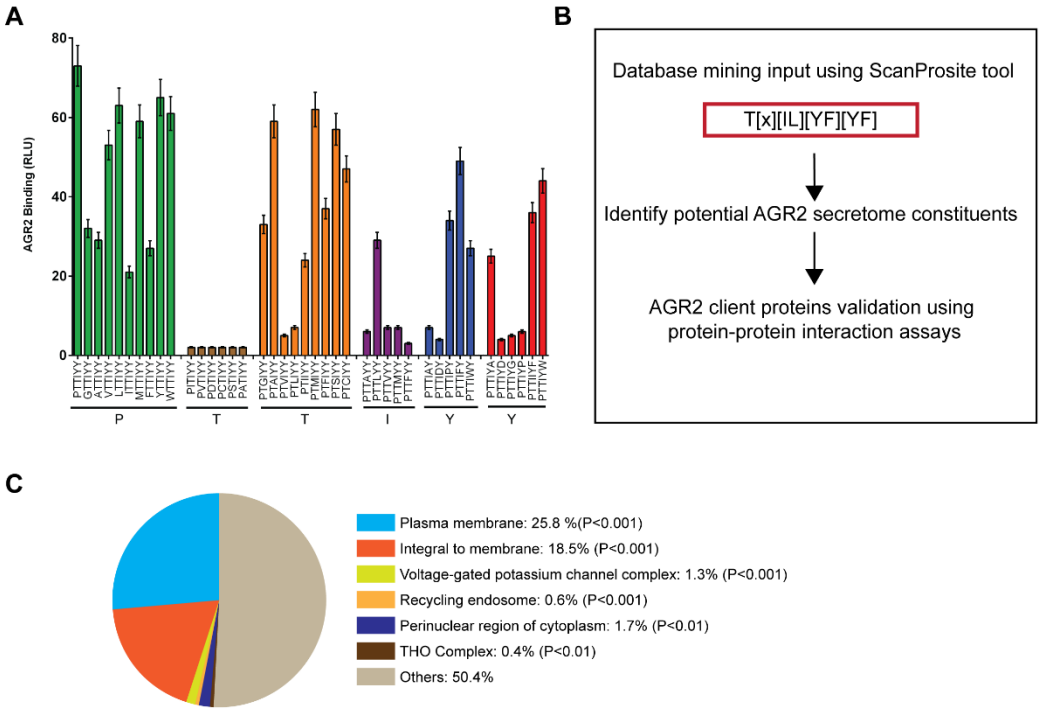
3.2.2. Developing a consensus peptide-binding motif for AGR2

AGR2 is an ER-resident molecular chaperone that plays a role in the maturation of cysteine-rich receptors such as MUC5 [104] and EGFR [83]. To our knowledge, AGR2 protein is relatively unique as a chaperone in possessing a sequence-specific peptide binding activity [209] and as discussed in 3.2.1.1. Therefore, it is vital to determine whether such a protein-protein interaction plays any role in binding to potential client proteins. If so, this would place a potential filtering step for client proteins with such a motif on AGR2-dependent processing. Previously, peptide A4 was shown to affinity purify monomeric or dimeric AGR2 protein and suggest the bioactivity of this peptide motif in crude lysates to AGR2 is relatively specific. A series of deletions and alanine scan mutagenesis of the 12-mer linear peptide A4 HLPTTIYYGPPG narrowed down the essential binding residues to only PTTIYY [209]. The originally optimised AGR2-binding peptide, PTTIYY, was subjected to mutagenesis in this study to produce a synthetic peptide library in which each of the six positions was mutated to a range of amino acids that are either charged, hydrophobic (bulky), and hydrophobic (small) (Figure 3.10A). AGR2 binding was measured by capturing biotinylated versions of the peptides onto streptavidin-coated plates and then assaying the extent of AGR2 binding. Relative to the positive control of AGR2 binding to the WT-peptide PTTIYY (Figure 3.10, far left bar), mutating each position resulted in the production of the consensus Tx[IL][YF][YF] (Figure 3.10B). For example, mutation of threonine at position 2 to any substitution abrogated activity (Figure 3.10A) suggesting a critical role for Threonine. Similarly, mutation Tyrosine at position 5 or position 6 was only tolerated by substitution to the bulky hydrophobic residues Phenylalanine and Tryptophan (Figure 3.10A). Mutation of Isoleucine at position 4 was only tolerated by a Leucine substitution (Figure 3.10A). By contrast, a

large set of amino acid substitutions were tolerated at Proline position 1 or Threonine position 3 (Figure 3.10A).

Using the optimised stringent consensus site Tx[IL][YF][YF], the human proteome was searched against the motif using ScanProsite[11] tool to identify potential human proteins that harbour the binding motif (Figure 3.10B). This database search resulted in 409 hits when splice variants or multiple isoforms of the same protein were excluded (Figure 3.10C-D). Interestingly out of 409 proteins, over 40% were predicted to be membrane associated (Figure 3.10C-E and Appendix 2). Among other hits were cilia-associated proteins, secreted proteins, protein-folding and transport-associated proteins, glycosylases, solute transporters and ion channels, and cancer-associated proteins. When gene enrichment tool was applied to the protein hits using FunRich [212], molecular functions related to membrane activity such as G-protein coupled receptor (GPCR) was among the highest (Figure 3.11A and Appendix 3). In the context of biological processes, proteins involved in transport, cell communication and signal transduction were enriched (Figure 3.11B). Signalling of GPCR and Olfactory were enriched in term of biological pathway (Figure 3.11C). This result suggests that AGR2 might involve in chaperoning lots of membrane class proteins and secreted proteins pathway. This is also consistent with previous studies which showed that AGR2 chaperone a number of secreted proteins such as MUC2 and MUC1 and also recently published paper showed binding to membrane bound protein EGFR. Furthermore, a large number of all the hits from the search were large, cysteine-rich proteins that undergo wide-ranging posttranslational modifications in accordance to proteins such as MUC2 and MUC1.

[11] http://prosite.expasy.org/scanprosite/



D

Gene names	Protein description	Motif
EPCAM	Epithelial cell adhesion molecule (EpCAM) (Adenocarcinoma-associated antigen)	TLIYY
TMEM67	Meckelin (Meckel syndrome type 3 protein) (Transmembrane protein 67)	TPIFY/TVLFF
TMEM63C	Calcium permeable stress-gated cation channel 1 (Transmembrane protein 63C)	TFLFF/TRLFF
MUC16	Mucin-16 (MUC-16) (Ovarian cancer-related tumor marker CA125)	TSIFF
FITM1	Fat storage-inducing transmembrane protein 1 (Fat-inducing protein 1)	TVIYF
KCNC3	Potassium voltage-gated channel subfamily C member 3 (KSHIIID)	TMIYY
OR2M4	Olfactory receptor 2M4 (HTPCRX18) (OST710)	TQIFF
SLC25A39	Solute carrier family 25 member 39	TAIYF
SMG7	Protein SMG7 (EST1-like protein C) (SMG-7 homolog) (hSMG-7)	TTIFY
TAS2R41	Taste receptor type 2 member 41 (T2R41) (Taste receptor type 2 member 59) (T2R59)	TLLFF
TRAM2	Translocating chain-associated membrane protein 2	TILFY
VPS33A	Vacuolar protein sorting-associated protein 33A (hVPS33A)	TLIFF
CACNA1B	Voltage-dependent N-type calcium channel subunit alpha-1B (Brain calcium channel III)	TTLYF
GPR133	Adhesion G-protein coupled receptor D1 (G-protein coupled receptor 133)	TILYY

Figure 3.10 Mining of the human proteome for proteins containing the AGR2 linear peptide consensus motif. (A) The PTTIYY hexapeptide was previously defined as a minimal peptide sequence that binds to AGR2 (19). A mutational scan library was synthesised containing a subset of amino acid substitutions at positions 1-6, from left PTTIYY. The substitutions included small hydrophobic (L, V, I, M, A, G), bulky hydrophobic (EpCAM, F, P), charged (D, C) and hydrophilic (S, T) of the linear peptide motif. The peptides sequences created are shown on the X-axis. Biotinylated peptides were bound to the streptavidin-coated solid phase and fixed amounts (1 µg) of AGR2 protein was added. AGR2 was detected using a secondary antibody and binding is measured in as RLU. The data revealed that amino acids at positions 2, 4, 5, and 6 are relatively fixed whilst changes at positions 1 and 3 can be relatively well tolerated. B). Schematic illustrating strategy to find novel AGR2 client proteins using linear peptide motif database mining. The AGR2 linear peptide consensus motif was used as input using a ScanProsite tool (http://prosite.expasy.org/scanprosite/) and the human proteome database was screened to identify proteins containing the motif. C). The scan resulted in 409 protein hits when splice variants were excluded (Supplementary Tables 1 and 2). The hits were scored as subcellular localisation using FunRich where the majority of the proteins found were membrane proteins (Appendix 3). A large proportion of the hits were membrane-related proteins, which foreshadows AGR2 function in receptor maturation, D). Representatives of possible AGR2 binding proteins are shown containing the consensus peptide-binding motif.

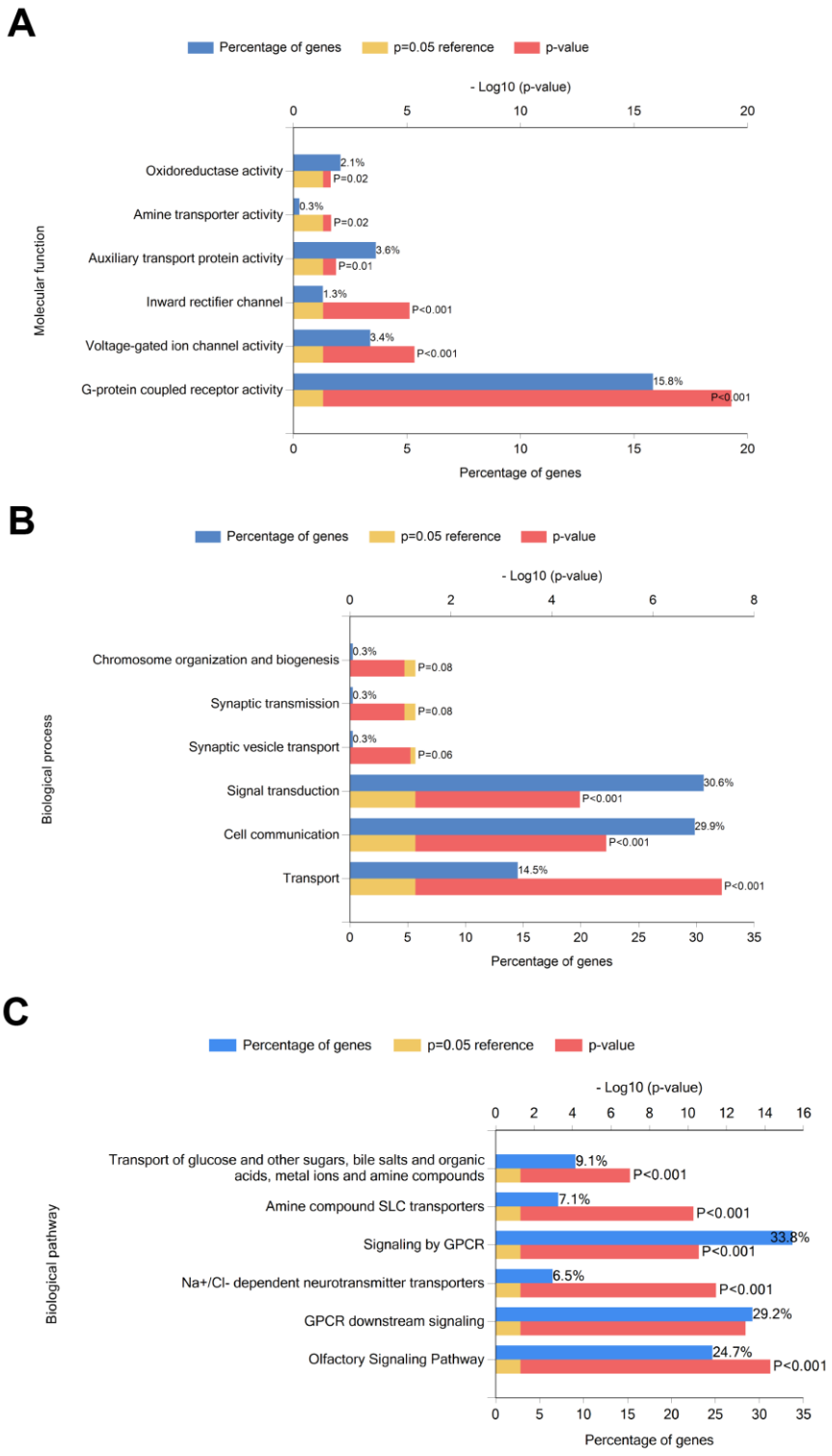


Figure 3.11 Functional enrichment analysis of AGR2 linear peptide motif hits using FunRich [212]. (A) Bar graph of molecular function overrepresented in AGR2 linear peptide motif hits. (B). Bar graph of biological processes overrepresented in AGR2 linear peptide motif hits (C) Bar graph of biological pathway overrepresented in AGR2 linear peptide motif hits. Proteins attributed to these enrichment were listed in Appendix 3.

As such, we focused on validating a key Tx[IL][YF][YF] motif-containing transmembrane receptor in this list that also possesses oncogenic activity to form a model protein to validate the role of this linear motif on AGR2 binding. This will help to understand the role of AGR2 in the secretory landscape of cancer by looking at functional consequences of interaction with protein for destined for secretory pathway. The receptor EpCAM fits these criteria as it contains a TLIYY motif in its extracellular stalk structure near the plasma membrane Figure 3.12A-B). Positions of the disulphide bonds in relation to the single pass transmembrane region and the TLIYY motif are shown in Figure 3.12C. Recent proteomics screen of 14 oesophageal adenocarcinomas versus two normal tissues biopsies by the Hupp lab identified EpCAM as the most independent validated "target" in human cancer as this protein was found to be overproduced proteins oesophageal adenocarcinoma [Personal Communication, 213]. EpCAM is a type I transmembrane 39–42 kDa glycoprotein that functions as a homophilic, epithelial-specific intercellular cell–adhesion molecule [214]. The 314 amino acid-long EpCAM protein comprises a large extracellular domain with an epidermal growth factor (EGF)-like domain and a putative thyroglobulin (TY) domain, a single transmembrane region, and a short (26 amino acids) cytoplasmic tail. The extracellular domain is sometimes referred to as EpEX, and the intracellular domain is sometimes referred to as EpICD. EpCAM has been shown to play role in cell-cell adhesion, cell motility, proliferation, survival, carcinogenesis and metastasis formation [215]. Overexpression of EpCAM in several types of cancer has been associated with poor prognosis [216]. EpCAM has been used as targeted cancer immunotherapy. For example, catumaxomab (Removab®) is a monoclonal bispecific trifunctional antibody which targeted EpCAM and T-cell receptor CD3 and this was approved to be used in the European Union in 2009 [217] for the treatment of patients with malignant ascites due to epithelial carcinomas. EpCAM was also used as circulating tumour cells (CTC) biomarker [214, 215, 218]. Altogether, this further favours EpCAM as a key putative AGR2 interacting protein and if so would shed the light the role of AGR2 in maturing membrane-bound proteins destined for secretory pathway.

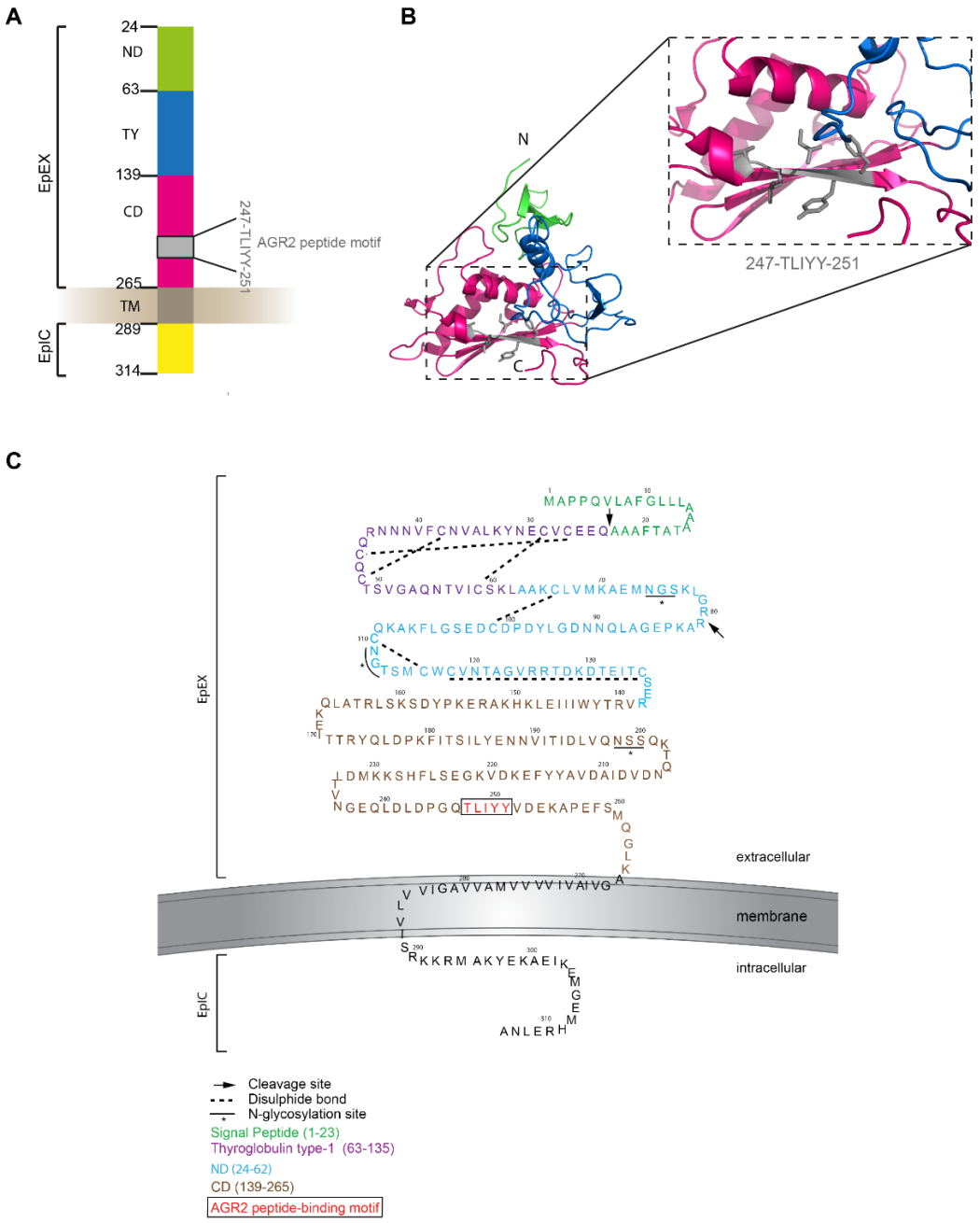


Figure 3.12 EpCAM as a candidate for AGR2 client proteins. (A) The secondary structure of EpCAM which consists of an N-domain (ND, green), Thyroglobulin type-1 domain (TY, blue) and C-domain (CD, dark pink) which altogether make up for extracellular part (EpEX), transmembrane domain (TM, grey) and intracellular part (EpIC, yellow). B). Three-dimensional cartoon representation of extracellular part of human EpCAM (PDB code: 4MZV) highlighting the AGR2 linear peptide motif at amino acid position Thr247 to Tyr251 9 (grey). Colour coding is the same as in (A). C). Schematic representation of full-length EpCAM protein sequence highlighting the i) extracellular part (EpEX) containing thyroglobulin type-1, N domain, C-domain, and AGR2 linear peptide motif ii) intramembrane iii) and intracellular part (EpIC). Arrows indicate cleavage site for signal peptide between position 23Ala and 24Gln by signal peptidase in the ER lumen and previously reported cleavage at Arg80 and Arg81. EpCAM contains three N-glycosylation sites (Asn74, Asn111 and Asn198) which are abolished in the crystal structure. EpCAM structure is stabilised by three closely spaced intramolecular disulphide bridges (Cys27-Cys46, Cys29-Cys59 and Cys38-Cys48).

3.2.3. Validating EpCAM as a specific AGR2 interacting protein

3.2.3.1. Solid-phase ELISA binding assay

To test whether AGR2 can interact with EpCAM, we chose to perform *in-vitro* solid phase binding assay which is a protein-protein interaction assay based on ELISA that was previously used as described in Figure 3.8. However, one of the binding protein is noncovalently coated by adsorption directly to the well surface of a 96-well polystyrene assay plate. The portion of the partner protein bound to the immobilised protein is quantified through antibody against the soluble partner protein using enzyme-linked immunosorbent assay (ELISA). This binary assay format ideally required a purified form of proteins. Since we have purified AGR2 successfully, we plan to purify EpCAM with the same affinity purification method.

3.2.3.2. Purification of EpCAM

EpCAM is a single-pass membrane protein which consists of an extracellular domain (242 amino acids) with an epidermal growth factor (EGF)- and thyroglobulin repeat-like domains, a single transmembrane domain (23 amino acids), and a short intracellular domain (26 amino acids). Since purifying the entire membrane proteins like EpCAM are challenging, only the extracellular part of the protein (first amino acid to 265 amino acid) was cloned. This extracellular part (termed EpEX) comprises the AGR2 peptide motif 'TLIYY' at position 247 to 251 that are sufficient for assaying protein-protein interaction in the solid-phase binding assay (Figure 3.13A-B). EpEX also was cloned without the N-terminal signal peptide which will be removed during EpCAM translation by a signal peptidase in the ER lumen. EpEX was first cloned into cleavable His-tag at N-terminal pEHISTEV vector which is expected to be at 32kDa. Following transformation into *E. coli* BL21 (DE3) bacteria and selection, isolated colonies were cultured in LB and induced with IPTG. Induction with IPTG at 37^0C for 3 hours yield low results (data not shown). The culture was then induced with IPTG at room temperature for 4 hours which give slightly better results. Coomasie staining of His-EpCAM when induced at room temperature for 4 hours and following his-tag affinity purification, showed bands in the elution fraction, however, the protein was not highly pure as the elution fractions contain multiple bands (Figure 3.13C). This

may be due to post-translational modification, specifically *N*-glycosylation of the protein which can occur in three potential *N*-linked glycosylation sites as highlighted in Figure 3.12C. However, western blot using EpCAM specific antibody showed that His-EPCAM is immunoreactive as it can be detected in the elution fractions with the correct expected size (Figure 3.13D). There was also a lower minor band that maybe corresponds to EpCAM cleavage when running under reductive SDS-PAGE conditions possibly at Arg-80/Arg-81 that can be detected by EpCAM antibody (antibody epitope at C-terminal of EpCAM) as described by Schnell and colleagues [214].

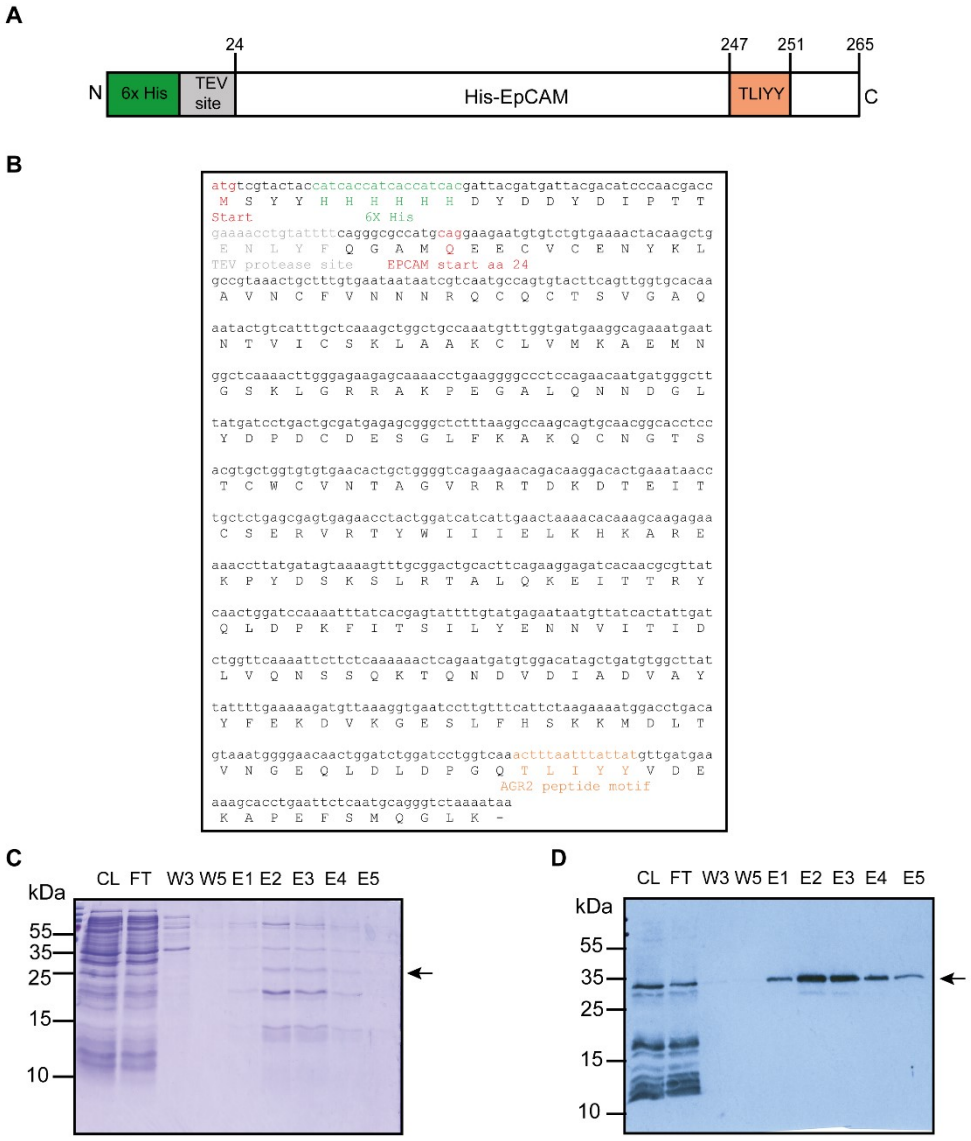


Figure 3.13 Purification of recombinant His-EpCAM. (A) Schematic representation of his-tagged EpCAM construct highlighting the extracellular domain (EpEX), the TEV cleavage site, and the TLIYY motif. (B) DNA and amino acid sequence of His-EpEX construct (C) Coomassie Blue staining of purified His-EpCAM showing multiple bands under reducing SDS-PAGE condition. (D) Western blot of purified His-EpCAM detecting His-EpCAM at 32kDa. CL= Crude lysate; FT= Flow Through; EpCAM=wash; E; Elution.

After having a poor EpCAM yield and purity upon affinity purification, we decided to use codon optimisation approach which can eliminate codon usage bias in certain host organisms such as bacteria. Codon optimisation has been proved to boost the levels of mRNA stability and productivity which in turn leads to higher yields of the protein [219]. We chose to chemically synthesise EpCAM and have a bacterial codon optimisation using GeneArt[12] gene synthesis. This technology combines gene optimisation where it calculates optimal DNA sequence needed to encode the protein of interest and codon optimisation where it uses optimal codon usage of the gene to the codon preferences of the selected expression organism. The software evaluates up to 50 factors that can compromise mRNA stability, such as extreme GC content, ribosomal binding sites, consensus and cryptic splice sites, repeats, and secondary structures. The codon optimisation algorithms have shown to increase five different classes of protein including membrane protein [219].

EpEx was chemically synthesised and designed to be cloned into pET151/D-TOPO, harbouring EpEx with N-terminal His-tag (Figure 3.14A). The clone was then transformed into *E. coli* BL-21 (DE3) and the bacterial culture was induced with IPTG for 4 hours at room temperature. We also decided to harvest bacterial cells expressing EpEx with urea lysis buffer as described in (Material and methods) to maximise extraction of EpEx from the cells. Protein purification of His-EpEx showed significantly increase in protein yield and purity as Coomasie-blue staining showed a major single band at the expected size of 32kDa and the protein was also overexpressed in the crude lysates (Figure 3.14C). The His-EpEx fraction was also functional and immunoreactive as it can be detected with specific monoclonal EpCAM antibody in western blot despite having undergone harsh lysis buffer during purification (Figure 3.14D). This proves that the gene optimisation in a specific host organism can result in better protein production.

[12] https://www.thermofisher.com/uk/en/home/life-science/cloning/gene-synthesis/geneart-gene-synthesis.html#

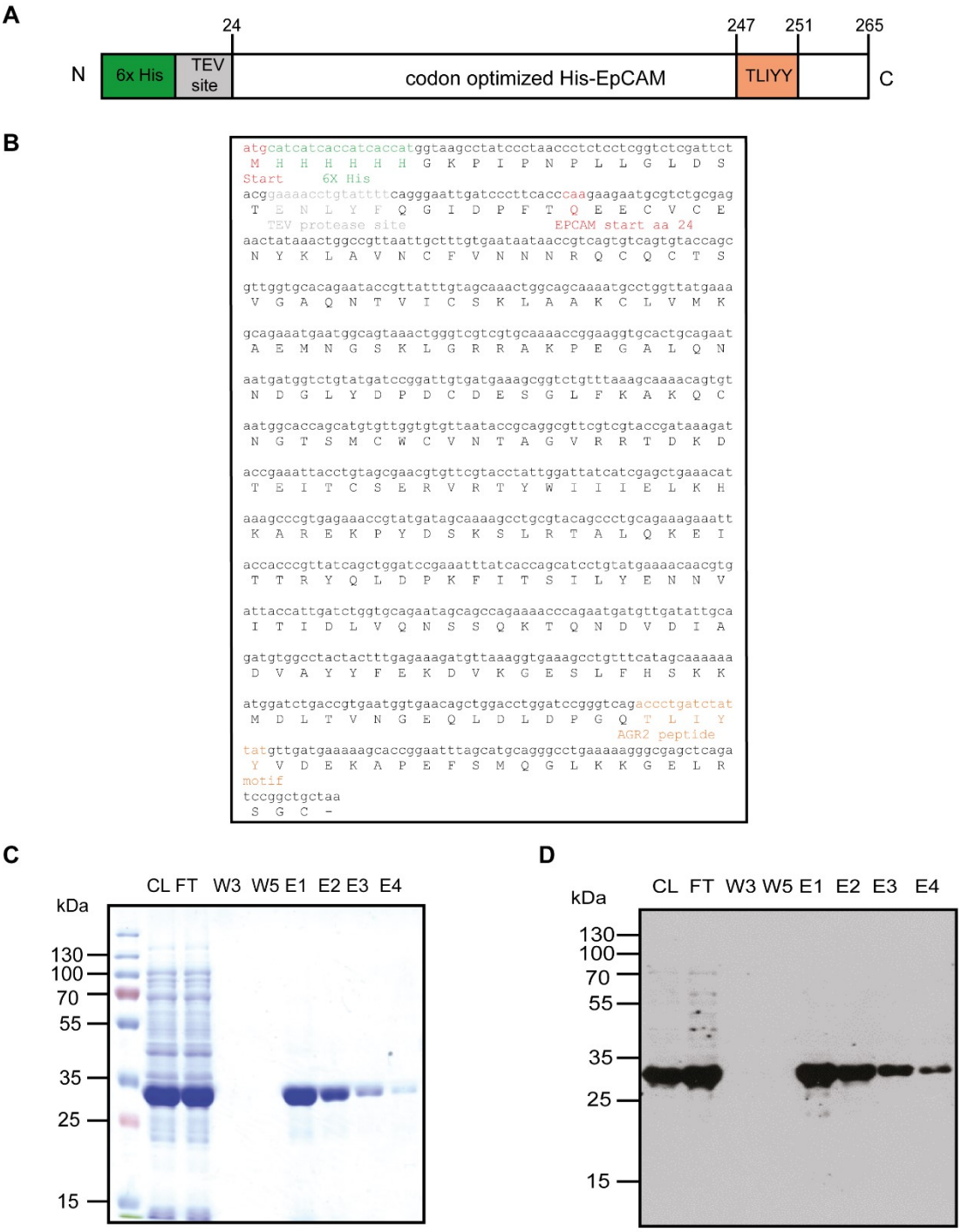


Figure 3.14 Purification of recombinant codon optimized His-EpCAM. (A) Schematic representation of chemically synthesized codon optimized his-tagged EpCAM construct highlighting the extracellular domain (EpEx), the TEV cleavage site, and the TLIYY motif. (B) DNA and amino acid sequence of the bacterial codon optimized His-EpEx construct (C) Coomassie Blue staining of purified His-EpCAM showing single major band at 32 kDa under reducing SDS-PAGE condition. (D) Western blot of purified His-EpEx detecting His-EpEx at 32kDa. CL= Crude lysate; FT= Flow Through; EpCAM=wash; E; Elution.

3.2.3.3. Solid-phase binding assay of EpCAM and AGR2

After having EpCAM and AGR2 purified we can then examine the binding of AGR2 and EpCAM in solid-phase ELISA binding assay. First, to make sure that the protein is properly coated on the 96-well polystyrene assay plate, His-AGR2 or His-EpCAM were coated overnight in carbonate buffer at 4^0C (Materials and methods) and the presence of the protein was directly detected by primary antibody followed by secondary antibody coupled peroxidase and quantification by chemiluminescence (Figure 3.15A-B, see first bar graph on the left and cartoon). The presence of immobilised His-AGR2 and His-EpCAM can be observed. Next, His-EpCAM was immobilised to the plate and AGR2 was added in the mobile phase and the binding was detected by the AGR2 polyclonal antibody (Figure 3.15A). Negative control wells were prepared without the addition of AGR2 primary antibody or His-AGR2 in the mobile phase to ensure that the binding is not due to unspecific binding of the primary antibody or secondary antibody respectively (Figure 3.15A). The results suggest that AGR2 was able to bind EpCAM in the solid phase. AGR2 binding to EpCAM also was observed when His-AGR2 was immobilised in the plate with EpCAM in the mobile phase followed by detection by the AGR2 polyclonal antibody (Figure 3.15B). However, the relative light signals were poorly detected compared to when EpCAM in the solid phase. We then decided to choose the format when EpCAM is coated to the solid phase for assaying binding to AGR2. The results showed that a dose-dependent titration of AGR2 in the mobile phase with increasing EpCAM protein on the solid phase demonstrated that the proteins form stable protein-protein interactions (Figure 3.15C). Polyclonal AGR2 antibody K47 was used throughout the thesis unless otherwise stated. The specificity of the antibody is validated and shown in Chapter 4 (Section 4.2.1 and 4.2.2).

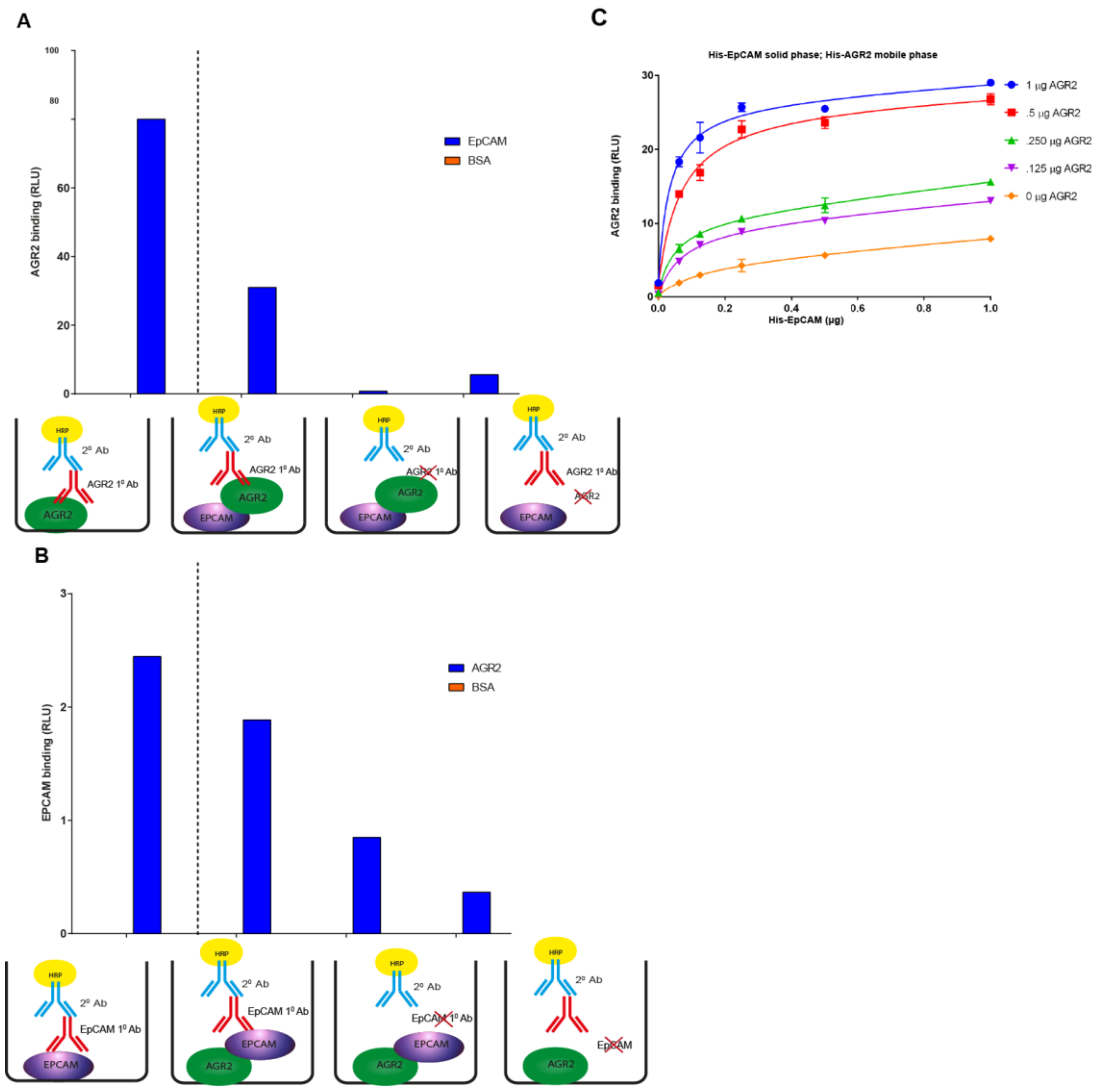


Figure 3.15 Solid phase ELISA binding assay to measure AGR2 binding to EpCAM protein. Solid phase ELISA binding assay was performed to detect the presence of His-AGR2 (A) or His-EpCAM (B) adsorbed on a solid phase. When soluble EpCAM (B) or AGR2 (A) is added to immobilise AGR2 or EpCAM respectively, binding activity is observed as a relative light unit. Controls were run to ensure the binding is not due to unspecific binding of the secondary or primary antibody to AGR2 as shown in the schematic. (C) Increasing amounts of EpCAM were immobilised on the surface of a microtiter plate (0-1 µg). AGR2 (0-1 µg) was titrated in the mobile phase and AGR2 binding to immobilised EpCAM was quantified using AGR2 specific antibody. The binding of AGR2 is plotted as the extent of protein-protein complex formation as RLU as a function of increasing protein in the mobile phase. The experiment was carried out in triplicate in four independent experiments, and data are representative of one experiment. The bars represent standard error of the mean (SEM).

Previously, AGR2 original phage display binding peptide was a 12-mer HLPTTIYYGPPG [209]. Truncation of the 12-mer from the N-terminal showed that the presence first Pro (P) residue was critical for AGR2 binding. However, this position was adaptable in its identity, where alanine scan showed that mutation of Pro (P) to Ala (A) did not significantly affect binding to AGR2. Similarly, truncation of the 12-mer from the C-terminal demonstrated that hydrophobic amino acid Tyr-Tyr (YY) is required for AGR2-binding, reducing the core binding motif to hexapeptide PTTIYY. When the 12-mer was mutagenized from tyrosine (Y) at the last position of hexapeptide to Ala (HLPTTIYAGPPG) dramatically reduce binding to AGR2 [209]. However, when the minimal peptide PTTIYY was mutagenized (Y to A) mutation at position 6 in this study can attenuate, but not abrogate, AGR2 binding to the synthetic hexapeptides (Figure 3.10A). Since EpCAM was found based on this motif with the sequence TLIYY, it is interesting to see if the mutation has the same effect as the AGR2 binding motif. His-EpCAM was then subjected to site-directed mutagenesis in which the Tyr-251 was mutated to alanine to create the TLIYA motif (Figure 3.16A, upper panel). His-EpCAMY^{251A} was then purified the same way as its wild-type counterpart. Purification of His-EpCAMY^{251A} showed relatively similar yield and purity and produced a major single band at 32kDa when stained with coomasie blue (Figure 3.16A, lower panel) and the proteins can also be detected with EpCAM antibody as analysed using western blot (Figure 3.16B). A titration of AGR2 against the wt-EpCAM or EpCAM^{Y251A} demonstrated that AGR2 binding was also attenuated on EpCAM^{Y251A} (Figure 3.16C). These data highlight at least one-binding site on EpCAM for AGR2 protein.

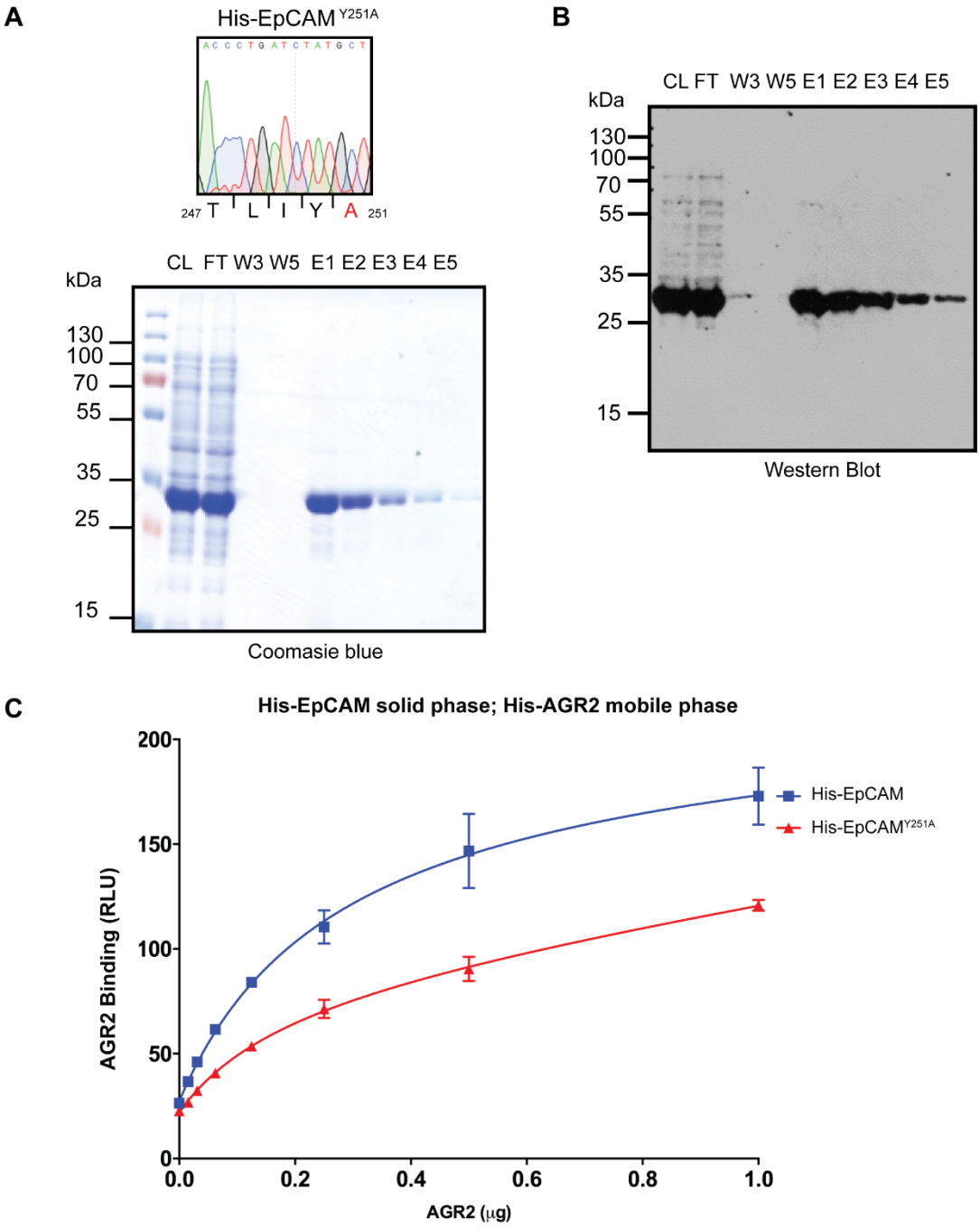


Figure 3.16 Mutagenesis, purification of His-EpCAM^{Y251A} and binding to His-AGR2. (A) Top panel - DNA sequencing chromatogram traces of EpCAMY251A. Bottom panel- Coomassie Blue staining of purified His- EpCAM^{Y251A} mutant protein showing a major band at 32 kDa under denaturing SDS-PAGE and western blot detected the same protein (B). CL= Crude lysate; FT= Flow Through; EpCAM=wash; E; Elution. (C) His-EpCAM or His-EpCAM^{Y251A} (1 µg) was immobilised onto the well surface of a microtiter plate. His-AGR2 WT (0-1 µg) was titrated in the mobile phase. AGR2 binding to immobilised EpCAM was quantified using AGR2 specific antibody. The binding is plotted as the extent of protein-protein complex formation as RLU as a function of increasing protein in the mobile phase.

3.2.3.4. Effect of mutation of the peptide-docking site on AGR2 in EpCAM protein binding.

The three docking site AGR2 mutants (D132A, P133A, and S134A) that exhibit neutral or attenuated binding to the synthetic peptide PTTIYY were evaluated in EpCAM protein binding to determine the effects of these mutations on the AGR2:EpCAM protein-protein interaction. The AGR2:EpCAM binding reaction revealed a gain-of-function of all three mutants relative to wt-AGR2, with $AGR2^{S134A}$ exhibiting the most elevated gain-of-function activity (Figure 3.17A). As AGR2 has a thioredoxin domain that can form covalent bonds with client receptors such as mucins [60], it can be expected that we evaluate the impact of reductant on its biochemical function. For example, the $AGR2^{C81S}$ mutant is as active as wt-AGR2 in stimulating ex-vivo cell growth [180]. Dithiothreitol (DTT) is a reducing agent which can be used to reduce the disulphide bond. The addition of DTT can determine whether disulphide bond contributes to the binding of AGR2 and EpCAM. The inclusion of DTT into reactions at the AGR2-binding stage (Figure 3.17A and B), surprisingly, exacerbated the differences between the mutants and wt AGR2, with the S134A followed by the D132A mutants exhibiting the most pronounced gain-of-function activity (Figure 3.17B and C).

We next staged this effect of reductant by including DTT only in the blocking stage, after EpCAM absorption, but before AGR2 binding (Figure 3.17D). This addresses whether this effect is due to an altered redox-imposed conformation of the substrate, EpCAM or the chaperone, AGR2. The incubation of DTT in the blocking step (Figure 3.17D, upper panel) is the stage at which the effect of the S134A mutant of AGR2 shows the most dramatic difference compared to wt-AGR2 (Figure 3.17D, lower panel). These data suggest that oxidised EpCAM, rather than AGR2, attenuates the impact of peptide docking site mutations on AGR2 functions. These data also suggest that the conformation of $AGR2^{S134A}$ is altered to drive it into a "docking-independent" activity towards a second, oxidation-sensitive binding site on EpCAM. For example, this alternate AGR2-binding site on EpCAM is apparently sensitive to its redox state.

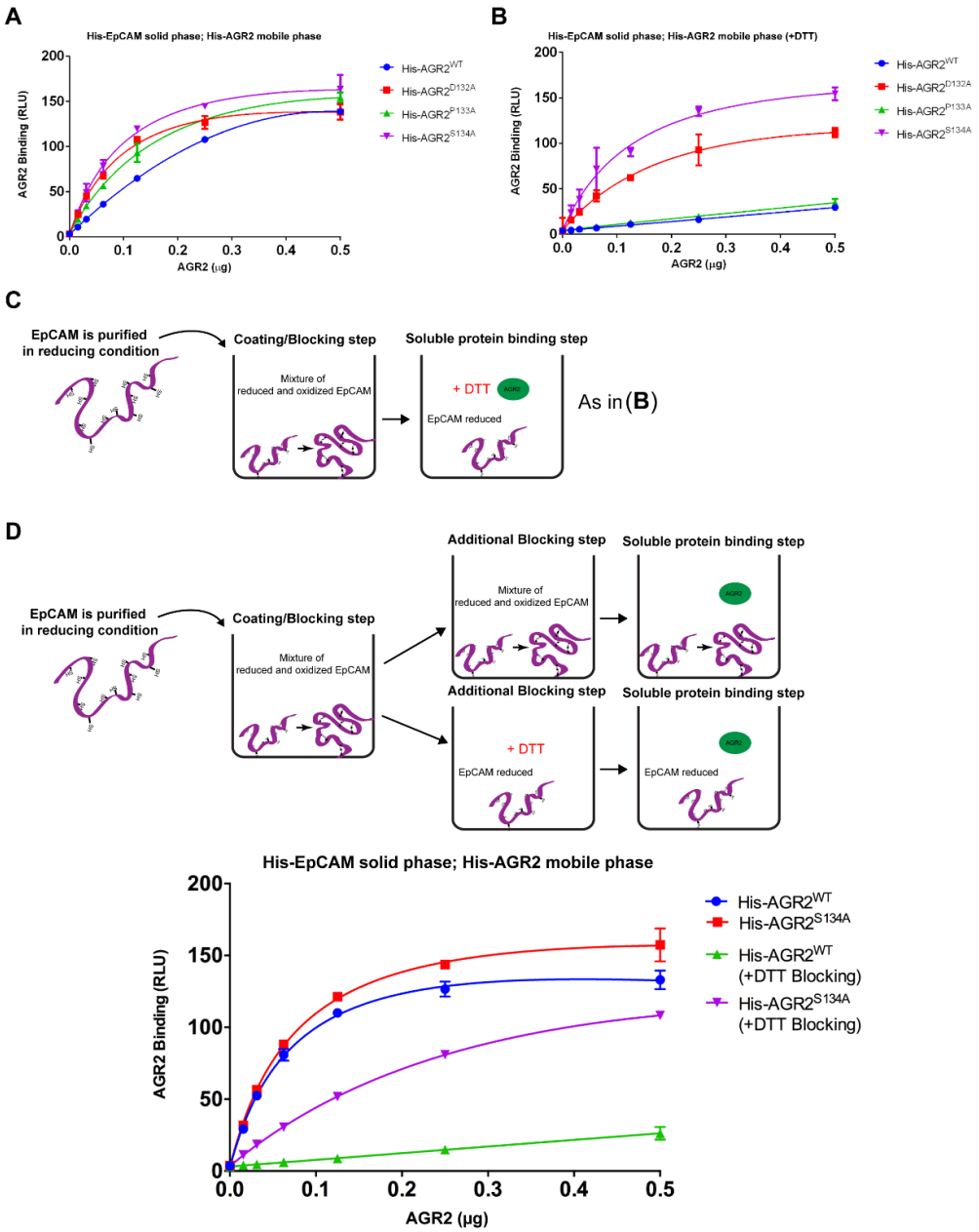


Figure 3.17 Effect of AGR2 peptide docking mutants binding to EpCAM. (A) His-EpCAM was immobilised onto the well surface of a microtiter plate. Titration of AGR2 WT and mutants (0-0.5 µg) were added in mobile phase and AGR2 binding without (A) and with 1mM DTT (B) to immobilised EpCAM was quantified using a specific AGR2 antibody. The binding is plotted as the extent of protein-protein complex formation in RLU as a function of increasing protein in the mobile phase. Reactions in (A) have no DTT included in the AGR2-binding reactions and (B) includes DTT at the stage of addition of AGR2 binding to reveal any effects of potential cysteine oxidation on protein-protein interactions (as highlighted in C where both AGR2 and EpCAM should be in the reduced state). (D) Staging the effects of DTT by including reductant in the blocking step to determine whether the DTT effect (from B) on the reaction is due to EpCAM substrate and not AGR2 itself. (E) A titration of wt-AGR2 and AGR2S134A in the presence or absence of reductant in the blocking stage. As in (A and B), the binding was plotted as the extent of protein-protein complex formation in RLU as a function of increasing protein in the mobile phase.

3.2.3.5. Evaluation of AGR2^{S134A} conformational change by hydrogen-deuterium exchange mass spectrometry

To evaluate whether AGR2^{S134A} does in fact, have an altered conformation, hydrogen-deuterium exchange mass spectrometry was used to probe its overall interaction with solvent. The wild-type and AGR2^{S134A} mutant proteins were diluted with D_2O and quenched from 30 seconds to 10,800 seconds post dilution. Relative to wt-AGR2, there was enhanced deuteration at several regions throughout the mutated AGR2^{S134A} protein (Figure 3.18 A-F for representative peptic fragments and Appendix 4). The most pronounced changes were observed at the dimer interface (aa 51-71) and at the peptide-docking site (aa 131-141) (Figure 3.18 G-I). For instance, the relatively selective and enhanced deuteration of the peptide-binding domain at the 30-second site point (Figure 3.18G) suggests the region is intrinsically more solvent exposed, although this does not prevent binding to the PTTIYY peptide. However, the pronounced exposure of the dimer interface at later time points (Figure 3.18H and I) to solvent suggests allosteric effects of the AGR2^{S134A} mutation impact on a monomer-dimer equilibrium that in turn expose the dimer motif to solvent. Indeed, at early time points after deuteration, when the protein is diluted 10 fold into D_2O, there is little difference in deuteration at the dimer interface. Thus, we would suggest that allosteric effects in the AGR2^{S134A} peptide-docking site could impact on monomer-dimer equilibrium. For example, the 10-fold dilution of AGR2 (at a starting concentration of 3 μM before dilution with D_2O and 0.3 uM after dilution) will take its final concentration lower than its published Kd of 8.8 μM [61]. Thus, such a dilution that is intrinsic to the deuterium exchange methodology will shift AGR2 into its monomeric state. Consequently, in the ELISA reaction measuring EpCAM binding, the wt-AGR2 and AGR2^{S134A} proteins are diluted to a final concentration from 0.01 to 0.2 μM where we would expect AGR2^{S134A} to be more conformationally altered and explaining in part its gain-of-function activity towards EpCAM. Together, our biochemical data define a dominant peptide-binding pocket on AGR2 and highlight that mutation of this motif is not localised but it can impact on heterologous protein-protein interaction such as EpCAM protein.

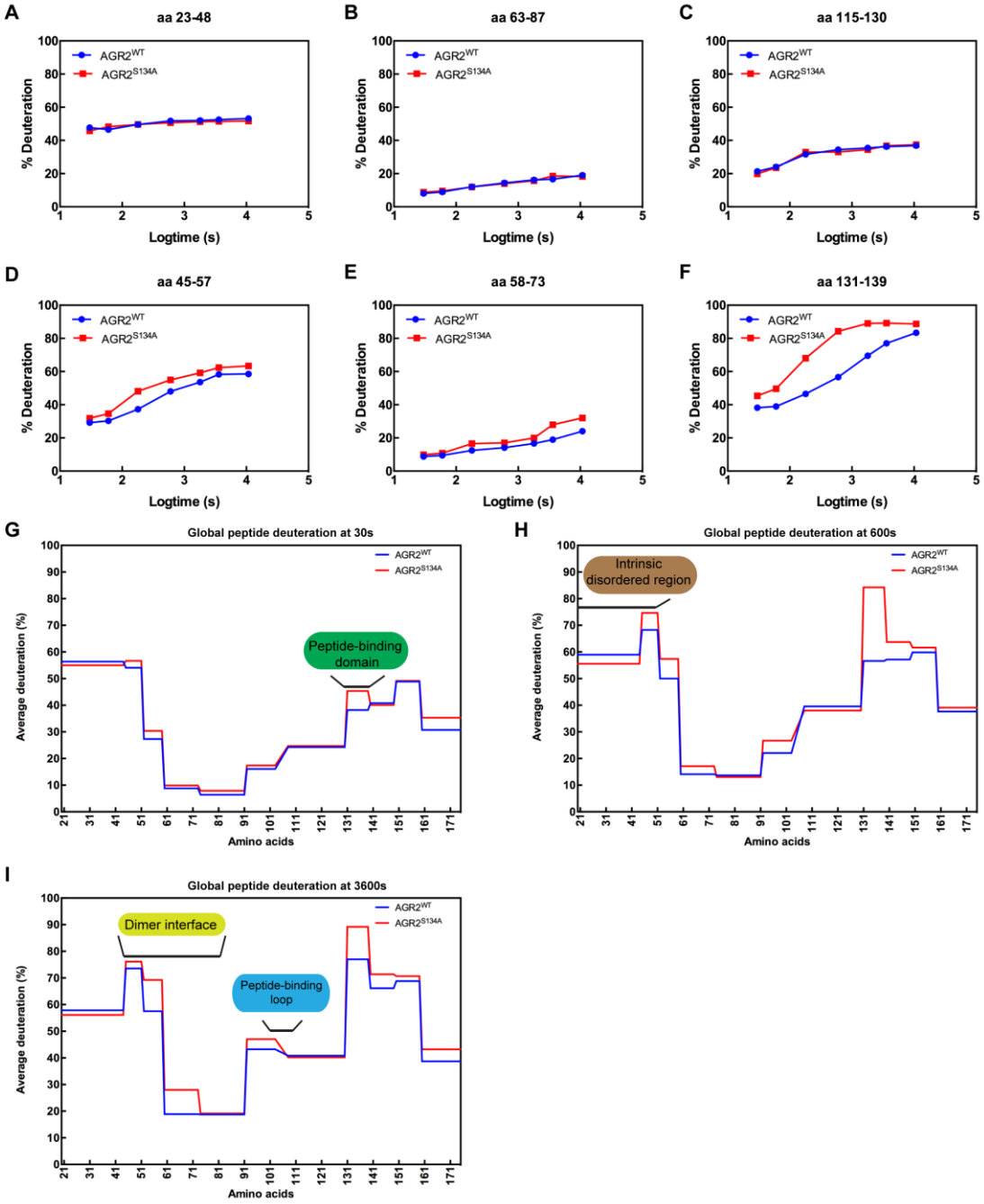


Figure 3.18 Measuring conformational changes in the gain-of-function mutant AGR2S134A using hydrogen-deuterium mass spectrometry. Wt-AGR2 or AGR2^{S134A} was deuterated over a 7-point time course from 30 seconds to 10,800 seconds followed by acidification, pepsinisation, and separation of fragments using mass spectrometry (Supplementary Figure 3) (A-C). Representative peptic ions of AGR2 protein that do not exhibit significant changes in deuteration between wt and mutant AGR2^{S134A} (D-F). Representative peptic ions of AGR2 protein that do exhibit significant changes in deuteration between wt and mutant AGR2^{S134A}. The data are plotted as % of deuteron exchange as a function of time (log10 in seconds; from 30, 60, 180, 600, 1800, 3600, and 10800). The deuterium exchange rates of individual peptides are summarised using the HDX exchange plots for (G) 30 seconds (H) 600 seconds, and (I) 3,600 seconds time course. In (G), we highlight that the most noticeable difference is enhanced deuteration at the "peptide binding domain", including aa 131-135. However, at elevated times of deuteration (H-I), there is apparent exposure of the dimerization domain and the peptide-binding loop to solvent, suggestive of a global conformational change induced by the S134A mutation.

3.3. Discussion

To understand the diverse biological functions of AGR2, the discovery of AGR2 interacting proteins is required to provide mechanistic insights. Yeast-two hybrid methods have identified potentially important AGR2-binding proteins. The first yeast-two hybrid identified the membrane receptor and pro-metastatic protein Dystroglycan/C4.4a as an AGR2 interactor [58] and that the newt homologue of AGR2 was shown to bind to the receptor Prod1 [54]. These data are consistent with the paradigm that AGR2 stimulates receptor maturation linked to growth or adhesion [94]. Another yeast-two hybrid screen revealed that AGR2 could bind not only to membrane receptors but many nuclear proteins such as RIP140 and Reptin, which might reflect a role for the form of AGR2 that escapes the endoplasmic reticulum through its non-canonical KDEL retention sequence [220]. The most well characterised AGR2 binding interactor is the AAA+ superfamily chaperone protein Reptin [105] [221] [97]. However, perhaps the most intriguing biochemical interaction function for AGR2 is in its intrinsic sequence-specific peptide binding activity [209]. In this chapter we exploit this specific "peptide-binding" function of AGR2 to propose a consensus-binding motif for its client proteins.

The majority of protein sequence information in higher eukaryotes is encoded by the "linear motif"; perhaps millions of such motifs are comprised of small stretches of amino acids that drive weak but specific protein-protein interactions. A powerful tool to define the linear motif repertoire for a protein is the peptide-phage combinatorial library [222]. This methodology has been used to define the peptide consensus motif for the peptide binding groove of the oncoprotein MDM2 [223]. This information stimulated the development of peptide mimetics that inhibit MDM2 *in-vivo* [224] [225].When AGR2 was subjected to a peptide-phage library screen, a pentapeptide docking motif emerged as TTIYY [209]. There were two distinct peptides selected in this study, but we have only focused on the one with the TTIYY motif. Synthetic peptides with this TTIYY motif could affinity purify AGR2 from cell lines (Figure 3.3) or human tissue biopsies with a high specificity [209] suggesting it could act as a highly specific recognition motif for AGR2 target proteins. This suggests

that the peptide could act as a highly specific recognition motif for AGR2 target proteins.

In this chapter, we apply hydrogen-deuterium exchange mass spectrometry to determine whether we can identify a specific binding site of the TTIYY containing peptide on AGR2. We identify a small region from aa 131-135 in AGR2 as the likely, primary binding site for TTIYY motif-containing peptides since it is here we see the most significant deuterium suppression after peptide binding (Figure 3.8 and Figure 3.9). Additional residues C-terminal to the 131-135 aa motif, including RADITGRYSNRL, also show deuterium suppression by the consensus site peptide A4, however, this region was not evaluated. We focused on aa 131-135 since this represents an exposed surface unstructured turn that might form a binding finger into hydrophobic TTIYY docking sites. To further validate this data, we found that upon mutating the motif, we could create a set of mutants with the loss of function on synthetic peptides (Figure 3.9), suggesting that this motif is the major peptide-docking site. By contrast, the same series of AGR2 mutant proteins exhibited an inverse gain-of-function activity on the client protein EpCAM (Figure 3.19) whose interpretation is complicated due to the oxidation-dependent resilience of EpCAM to this AGR2 mutant series (Figure 3.17). Nevertheless, the ability to create a loss or gain-of-function mutations in this region suggests that it not only provides a major primary interaction site but that the conformation of this region can allosterically affect other regions that impact on monomer-dimer equilibrium. The enhanced binding of EpCAM to this S134A mutant, in particular, suggests that AGR2 might have a secondary binding site on EpCAM that is stabilised by the altered conformation of the S134A mutant. This potential secondary site appears to play only a minor role on wt-AGR2 protein since the $EpCAM^{Y251A}$ mutation attenuates wt-AGR2 (Figure 3.16).

The human proteome was scanned for proteins containing this motif in order to produce a set of potential AGR2-client proteins and there was a relative enrichment in transmembrane proteins (Figure 3.10). There is a degree of specificity in this peptide motif, as a prior selection of peptides from phage libraries to p300 identified PxxP motif proteins that predominate in transcription factors like SMAD4 and p53 [226].

However, none of the AGR2 client proteins identified in previous studies overlap with our method of identification, suggesting that our method is exclusive in determining novel AGR2 client proteins. This provides a complementing approach to the yeast-two hybrid and immunoprecipitation methodology to identify new interacting proteins by exploiting the linear peptide motif as a common type of protein-protein interaction [196].Most of the membrane proteins identified from the database mining possess the motif at their extracellular domain. Nascent protein chains that translocate into the ER associate with specific ER luminal proteins, which in turn ensures proper folding and post-translational modifications. Since AGR2 localises in the ER lumen, AGR2 might be one of the key proteins in aiding protein maturation suggesting that protein-protein interaction may occur in the ER.

EpCAM was chosen as the target of choice and its paralogue (TACD2) also shows identity in this region (Figure 3.11). In this chapter we only showed using ELISA that purified AGR2 and EpCAM can interact with each other. It is also noteworthy in the future to perform Circular Dichroism (CD) spectroscopy studies of EpCAM and its tyrosine mutant in binding to AGR2 in the future. CD is a biophysical quantitative technique in which the CD of molecules is measured over a range of wavelengths. This method can determine whether there are changes in the structural integrity of proteins when they interact by looking at the changes in CD spectra which gives evidence that there is a direct binding. However, CD requires relatively large amounts of proteins for such biophysical studies which could be optimised by purification in insect cells. The consensus motif within EpCAM forms a β-strand within the main α-β fold in the EpCAM structure defined by PDB (4MZV) and this located in the extracellular region. It has been noted that linear motifs can reside within intrinsically disordered domains or they can reside within structural domains [206]. In the case of the linear motif within disordered regions, it is known that they can acquire distinct secondary structures depending upon the nature of the protein-interaction. If the linear motif resides within a structural domain, the protein-protein interaction presumably requires an alternate conformation at the target site. Using EpCAM as a model, we can speculate that AGR2 might interact with an unfolded version of EPCAM as the receptor is being assembled in the ER. This is also consistent with our

in-vitro protein-protein interaction assay using ELISA in which we used non-glycosylated EpCAM. This is because we used bacterial host (*E. coli*) which lacks an N-glycosylation system for our protein purification. We chose this purification as it is easy to make rather than using mammalian insect cells, and perhaps since AGR2 is also purified from the bacterial host system. Based on ELISA results, AGR2 can bind to non-glycosylated EpCAM which suggests that the interaction occurs in the ER prior to glycosylation. However it is also possible that carbohydrate attached to glycosylated EpCAM is involved in the interaction since EpCAM possesses three N-linked glycosylation sites. The presence of carbohydrate can modify its tertiary structure or its presence obscures parts of the protein to which AGR2 protein would bind in its absence. In addition EpCAM was shown to be hyperglycosylated in carcinoma tissue as compared with autologous normal epithelial, and glycosylated EpCAM demonstrated a degree of tissue specificity [227] and was shown to be crucial for tumour stability [228]. Therefore, it would be appealing to establish whether EpCAM glycosylation has any effect in binding to pro-oncogenic AGR2 or whether AGR2 has any role in the glycosylation of EpCAM in the future. In addition, EpCAM can be cleaved producing intracellular domain (EpICD) and extracellular part (EpEx) by an intramembrane proteolysis [229, 230]. The EpEx fragment has been found in microvesicles (e.g., exosomes) which represent potential signals that can migrate from cell to cell inside of the same organ or circulate into the bloodstream to distant organs. Since EpEx fragment contains AGR2 peptide motif and the fact that AGR2 can be secreted, it would be interesting to study the relationship of the extracellular role of AGR2 and EpCAM in tumour signalling. Intriguingly, a recent study described the unexpected role of an extracellular AGR2 (eAGR2) as a regulator of epithelial tumorigenesis and morphogenesis [180] suggesting another role of AGR2 in cancer development. Alternatively, AGR2 might interact with the receptor as receptor conformation is altered especially as the peptide-docking site is near single-pass transmembrane domain. We can also speculate that since the docking site is in the C-terminus of EpCAM N-terminal to the single pass transmembrane domain, if AGR2 binds to this inside the ER, the motif could impact upon how the protein folds as the C-terminus is synthesised into the ER membrane. The docking site might facilitate a lower off-rate for AGR2 on EpCAM inside the ER to allow the correct formation of

key disulphide bonds. Alternatively, the docking site might facilitate AGR2-mediated transport of EpCAM cargo as it is transported to the plasma membrane. A key observation is that in cells, the EpCAM^{Y251A} mutant protein fails to reach its normal plasma membrane destination, but is present in the cytosol or nuclear membrane (will be discussed in Chapter 4). There is a redox component to this docking site *in-vitro* (Figure 3.18), suggesting that disulphide bridge formation in the EpCAM protein can impact on how AGR2 binds in cells, although cell-based disulphide shuttling assays might be more difficult to reconstitute in cell systems. Future cell-based assay developments can impact on understanding how the EpCAM docking site (aa 247-251) facilitates its maturation or trafficking and how the AGR2-peptide binding motif impacts upon client protein maturation. For example, it would be interesting to generate a gain-of-function AGR2^{S134A} or cysteine-mutated EpCAM cell line using CRISPR/Cas9 gene editing cancer to measure changes influx of EpCAM receptor maturation as a result of such mutations.

To conclude, we have provided a methodology to exploit the linear peptide motif as a tool to discover new protein-protein interactions for a target protein This includes mapping a specific site on AGR2 protein implicated in sequence-specific peptide binding and suggest a function for this motif in client proteins is to ensure proper maturation for its final destination. Transmembrane proteins are proposed to be the dominant class of proteins with such a motif and it will be interesting to stratify such "PTTIYY" motif containing proteins for AGR2-dependent maturation. Although these data provide an intracellular function for AGR2, it is interesting to note that on EpCAM, the AGR2 docking site is extracellular (Figure 3.12). Given the recent extracellular role for AGR2 in cancer cell growth [180], these data also suggest that AGR2 might function as an extracellular ligand to impact on EpCAM signalling in cells. Most of the work in this chapter utilises *in-vitro* biochemical approaches and there is no evidence whether AGR2 and EpCAM are indeed interacted with each other. Therefore, it is important to find such evidence *in-vivo*. This also serves as proof that our method of discovering client proteins is valid. In the next Chapter, we discuss protein-protein interaction assays in cells and utilises some of the techniques to further validate our AGR2-EpCAM interaction.

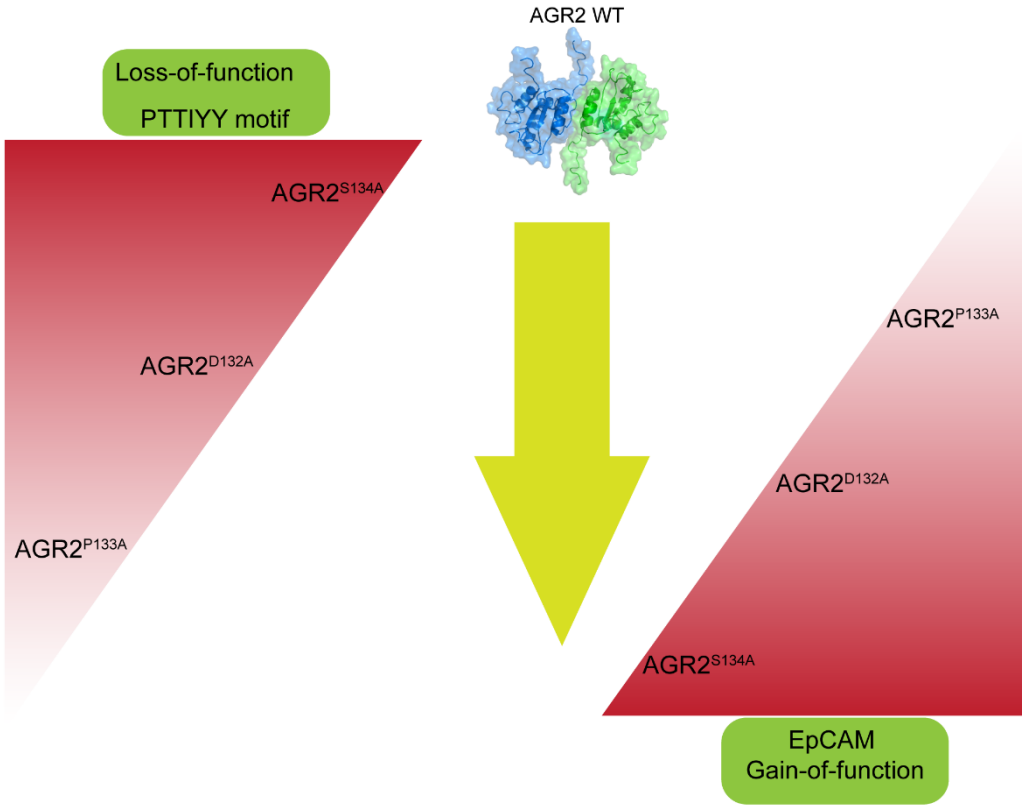


Figure 3.19 . Summary of the biochemical properties of the peptide-docking site mutations in AGR2 produced based on hydrogen-deuterium exchange mapping. Based on the hydrogen-deuterium exchange mapping data (Figure 4 and 5), we focused on creating three alanine substitutions mutations in the VDPSL loop motif, from aa 131-135, residing between two β-sheets (Figure 6). The mutant proteins exhibit inverse trends in their specific activity. Consensus peptide binding reactions demonstrate that wt-AGR2=S134A>D132A>P133A with two mutants showing a loss-of-function. Whilst in EpCAM binding, S134A>D132A>P133A=wt-AGR2 with two mutants showing a gain-of-function. These data suggest that although mutating some amino acids in the VDPSL motif can impact on specific peptide binding, the global conformation changed induced by loop mutation (for example in the gain-of-function S134A mutation (Figure 11) results in binding to a distinct site on the EpCAM molecule. These data thus suggest that one purpose of the VDPSL motif is to not only drive specific-peptide binding by AGR2 but to allosterically constrain the conformational dynamics (or monomer-dimer equilibrium) of AGR2 so as to minimise its "binding to other sites" on its client protein

CHAPTER 4: Cell-Based Analysis of AGR2 and EpCAM Interaction

4.1. Introduction

4.1.1. Protein-protein interaction networks

The completion of the human genome project has provided an immense amount of information that aimed to generate a comprehensive picture of all relevant genotypic variations in the human population. For example, many polymorphisms associated with disease have been linked to Mendelian disorders [231]and over one hundred thousand alleles have been implicated in at least one disorder [232]. In addition, a large number of susceptibility loci have been identified for hundreds of complex traits [233]. Despite the massive information of genomics sequences, the connection between genotype and phenotype is still not fully understood [234], especially for multifaceted trait loci and cancer-associated mutations. Even if the phenotype is known, the mechanism of how mutation perturbs the function of the corresponding gene or gene product is not fully understood. To understand the genotype-phenotype relationship, functions and context must be assigned to large numbers of genotypic changes.

Cellular systems are connected through interactions among genes and gene products, or interactome networks, which represent most cellular functions [234]. Thus, to understand a comprehensive understanding of genotype-phenotype relationships in human, mechanistic descriptions of how interactome networks are distressed as a result of genomic susceptibilities are necessary. This may be achieved by building high quality and extensive genome and proteome-scale maps of macromolecular interactions such as protein-protein interactions (PPIs), protein-nucleic acid interactions, and post-translational modifiers and their targets.

Initial analysis of human genome estimated about 30,000–40,000 protein-coding genes existed in the human genome [235] and a subset of 10 000 proteins can be produced by the cell at any given time (termed proteome) [109]. This number could be underestimated due to alternative splicing that could create a larger number of

protein products. The functions of numerous of these proteins are not known to understand the dynamics of the human organism.

Protein-protein interactions (PPIs) are intrinsic to virtually every cellular process. Proteins are products of intertwined DNA and RNA regulations that facilitate most biological processes in a cell, including gene expression, cell growth, proliferation, nutrient uptake, morphology, motility, intercellular communication and apoptosis. Cells can respond to a numerous of external stimuli, which in turn dictates the expression or activation of specified proteins to perform the biological tasks. In addition, cells are varied, and many proteins are expressed in a cell type-dependent manner. Hence, these basic properties of proteins suggest a complexity that can be difficult to investigate, especially when trying to understand protein function in broader biological context.

Conventionally, protein function analyses were mainly focused on single isolated species [236]. However, this is not the case. It has been revealed that over 80% of proteins operate in complexes for proper function [109]. Therefore, the study of PPIs is imperative to fully understand protein function within the cell. For example, the functionality or behaviour of unknown proteins can be foretold on the evidence of their interaction with a known protein function. Phizicky and Fields describe several important outcomes of PPIs [110] and these include; (i) altering the kinetic properties of proteins; (ii) permitting substrate channeling; (iii) creating new binding site for small effector molecules;(iv) inactivating or suppressing a protein; (v) changing the specificity of a protein for its substrate through interaction with different binding partners; (vi) or providing an upstream or downstream regulatory role.

One of the advantages of studying protein-protein interaction is that it led to the identification of drug targets [237]. Proteins can involve a larger number of interactions built up by highly connected nodes (hubs) which can include families of enzymes, transcription factors, and intrinsically disordered proteins and many poorly connected nodes [238, 239]. Each node receives inputs and produces one or more specific outputs. This suggests that the latitude of PPIs regulation is large and may

represent more diverse processes. To accurately understand their importance in the cell, various interactions of a particular protein and the outcome of the interactions need to be taken into account.

The technology advancement has been driven by the high-throughput experimental methods, which has made the identification of proteins a relatively simple task. This high-throughput experimental approach includes mass spectrometry, yeast two-hybrid system, combinatorial phage-display and protein chip technology. Mass spectrometry, for example, can be used not only to identify individual proteins but also to characterise protein complex [240]. However, the massive discovery of these newly protein binders has not been fully validated. Conceivably, in recent years, computational analysis (*in-silico* approach) has gradually becoming popular to predict client proteins and to understand the function of unknown proteins[241].

4.1.2. AGR2 Protein-protein interaction networks.

Yeast-two hybrid (Y2H) and co-immunoprecipitation, two commonly used assays to study protein-protein interaction, have largely driven the comprehensive discovery of AGR2 binding proteins. For example, MUC2, a large, cysteine-rich glycoprotein that forms the protective mucus gel lining the intestine, was found by immunoprecipitation with AGR2, suggesting AGR2 involvement in mucin processing [60]. Studies using Y2H showed that human AGR2 can interact with the metastasis-associated GPI-anchored C4.4a protein and extracellular alpha-dystroglycan (DAG-1) [58] and membrane receptor Prod1 [54]. These data showed that AGR2 plays a pivotal role in the secretion of the receptor protein that is linked to adhesion and cell growth. The majority of studies only focused on the direct interaction of AGR2 and its client proteins, without fully interrogating the molecular mechanism of the interaction. It is important to note that Y2H is prone to error as numerous false positive discovery have been pointed out, thus detailed validation of partners is necessary to demonstrate biological significance [242]. Perhaps, there are many proteins identified as an AGR2 binding partner from Y2H studies but only a few were thoroughly validated [94]. Reptin is the AAA+ ATPase protein and the only AGR2 client protein found in the Y2H screen that has been thoroughly studied [105]. The AGR2:Reptin interaction is

based on ATP-de-stabilising manner, was mapped to a peptide interface comprising the proposed substrate-binding loop on AGR2, within amino acids F104 to Y111, that is shared with a binding loop of ERP18. Amino acid mutagenesis at that positions attenuates Reptin binding to the AGR2. Later it was shown that Reptin participated in the cytosolic regulatory protein-folding complexes through the chaperone-dependent R2TP activity [106] and in cilia complex assembly at the plasma membrane [107] that may drag AGR2 involvement.

In this study, we exploited the high-affinity AGR2 binding peptide from previous phage-peptide library data to expand potential AGR2 client proteins. We further characterised the binding peptide, in which we mapped the peptide docking site using hydrogen-deuterium exchange mass spectrometry, suggesting aa 131-135 on AGR2 is the primary peptide docking site (Chapter 3). We then developed a consensus site (TxIL[YF][YF]) for its activity and performed human proteome database mining using to identify potential cellular proteins that harbour this motif as potential AGR2 client proteins. Transmembrane/receptor proteins were identified as a dominant type of proteins with potential AGR2 docking sites. These findings are consistent with the concept that AGR2 can mediate secretion of receptors linked to adhesion and cell growth[94, 116]. We validated one key protein-protein interaction, represented by AGR2 and the oncogenic receptor EpCAM (containing TLIYY motif) and suggested one function for this sequence-specific peptide binding is to drive an interaction with proteins destined for the secretory pathway.

4.1.3. Aims of Chapter

In the previous chapter, a membrane-bound EpCAM protein was validated as an AGR2 interacting partner using *in-vitro* biochemical cell-free systems. This technique such as an ELISA is rather a simple and direct method for the analysis of protein-protein interactions. However, the dynamics of protein-protein interaction in cellular context might be more complex. Therefore, in this chapter, we aim to find evidence that AGR2 and EpCAM interact in cell system and if so, we aim to further characterise the interaction. This will help us understand the molecular role of AGR2 in trafficking a protein destined for secretory pathway and find a mechanism of such interaction. A panel of cancer cell lines was screened to establish cell lines that express or do not express AGR2 and/or EpCAM to serve as cancer models. Cell-based PPIs methods such as an *in-situ* proximity ligation assay and colocalisation of fluorescently-tagged AGR2 and EpCAM using confocal microscopy were performed. We also aim to exploit the FLO-1 cells (AGR2-/EpCAM-) in a reconstitution studies. The FLO-1 cells were used to reconstitute AGR2 and EpCAM expression to assess the localisation of each protein, and investigate how an ER-resident protein AGR2 can regulate and determine EpCAM expression and localisation. The final aim of this chapter was to assess the expression of AGR2 and EpCAM in tumour clinical samples and find evidence that two proteins are co-expressed using immunohistochemistry.

4.2. Results

4.2.1. Identification of AGR2 expressing cell lines as cancer models

Human cell lines derived from tumour biopsies are the most frequently employed models in cancer research and their use has facilitated the understanding of cancer biology. Cell lines offer the advantages such as easy to maintain and propagated, relatively inexpensive, and accessible to high-throughput testing of therapeutic agents [243]. Data generated from cell lines can then be extrapolated to *in-vivo* human and provide models for drug testing and translational study. In fact, AGR2 was first identified in mammalian studies in oestrogen receptor-positive breast cancer cells [126]. Therefore, it is important in this study to establish a cell line for proper expression of the specific protein (i.e whether the cell line expresses a target protein where its pathway is intact or vice versa) in order to fundamentally study the cellular pathways and to unveil critical genes especially those relating to its interactome in cancer.

A panel of cell lines containing breast cancer cell line MCF-7 and oesophageal adenocarcinoma; OE19, OE33 and FLO-1 were tested to assess the basal expression level (if any) of AGR2 and its putative binding partner EpCAM using specific AGR2 and EpCAM antibodies in western blot (Figure 4.1A). Western blot using rabbit polyclonal AGR2 K47 (Moravian Biotechnology) exhibited that MCF-7, OE33 and OE19 expressed AGR2 while FLO-1 cells did not express detectable expression of AGR2 (Figure 4.1A). OE19 relatively showed the highest expression of AGR2 followed by OE33 and MCF-7. Western blot analysis using EpCAM mouse monoclonal antibody (Calbiochem), EpCAM showed a similar pattern of expression in the cancer panels, where MCF-7, OE33 and OE19 are positive for EpCAM (Figure 4.1A). Interestingly, FLO-1 cells were also absent for EpCAM and this can serve a very good negative control or can be used for reconstitution complex studies for manipulating AGR2 and EpCAM protein-protein interaction assays in cells. Mouse monoclonal p53 (DO-1) antibody was also used on the same blot to examine p53 status in relation to AGR2 and EpCAM expression (Figure 4.1A). All the oesophageal adenocarcinoma cell lines contain mutated p53 while MCF-7 breast cancer cell lines have wild-type p53 [244, 245].

The specificity of AGR2 antibody was validated using short interfering RNA (siRNA) that targets AGR2. A set of siRNA to AGR2 (FlexiTube GeneSolution GS10551, Qiagen) and siRNA control was acquired (AllStars Negative Controls FlexiTube siRNA, 1027280, Qiagen) and transfected into a panel of cell line (Figure 4.1B). After incubation of 72 hours, the cells were analysed using western blot using AGR2 K47 antibody. AGR2 can be knocked down using siRNA 1 and 2 and to some extent siRNA 4 in MCF-7 and OE-33, where most knockdown was seen in MCF-7 compared to control siRNA or non-transfected controls (Figure 4.1B). However, AGR2 cannot be knockdown in OE-19 using the same siRNA set. As shown in western blot, OE19 expressed a high amount of AGR2 which might need a higher amount of siRNA and longer incubation to bypass the positive steady-state regulation of AGR2(Figure 4.1B). This data altogether, validated the use of AGR2 rabbit polyclonal K47 and band correspond to western blot was AGR2-specific.

We next examined the localisation of AGR2 and its client protein EpCAM in MCF-7 using immunofluorescence (Figure 4.1 C-D). MCF7 cells lines were stained with the same antibody anti-AGR2 K47 polyclonal and anti-EpCAM monoclonal antibodies. Using confocal microscopy, the localisation of endogenous AGR2 showed diffuse distribution and observe as punctate in the cytoplasm or specifically in the ER as previously shown [Fourtouna et al., 2009] (Figure 4.1 C). EpCAM was seen to localise mainly in the cell membrane and partially in the cytoplasm which consistent of a typical membrane protein localisation (Figure 4.1 D). These results served further evidence that the AGR2 and EpCAM antibodies were specific.

4.2.2. AGR2 and EpCAM interaction using *in situ* proximity ligation assay

Proximity ligation assay (PLA) is a recent development of detecting protein-protein interactions (Figure 4.2). It was first described by Fredriksson and colleagues in 2002 [246] and modified by Söderberg and colleagues et al. [247]. The PLA technique utilises simultaneous recognition of oligonucleotide labelled antibodies binding in close proximity (30-40 nm apart) by pairs of affinity probes of the same

protein or two proteins in a complex. Thus, the advantage of the proximity ligation assay is that it can detect an authentic endogenous protein-protein complex *in situ* and it does not rely on transfected or non-physiologically-tagged protein vectors. The assay can also be used to quantify protein and modification of proteins in fixed cells and tissue samples. Two primary antibodies raised in different species (such as one mouse and one rabbit antibody) are used to detect two target antigens (Figure 4.2). Two additional species-specific secondary antibodies are then introduced. The secondary antibodies are attached to unique DNA strands that serve as a template for hybridization of added oligonucleotides. If the two proteins are in close proximity (30-40nm) the oligonucleotides are ligated by ligase, leading to the formation of a circular, single-stranded DNA molecule. The DNA circles, in turn, provide templates for rolling circle amplification (RCA) by DNA polymerase. The addition of fluorescent probes then allows individual interacting pairs of protein molecules to be visualised with a fluorescent microscope. The main disadvantage of PLA is cross-reactivity and nonspecific adsorption, which sets the limit for the selectivity of such assays. Another drawback of PLA is that the concentration of proximity probes must be kept minimal in order to prevent unwanted background signals originating from accidental proximity of non-bound probes. Furthermore, since PLA can permit identification of pairs of antibodies located within less than approximately 40 nm which is higher than standard light microscopy (approximately 200nm), this resolution is still lower than that of FRET which detects molecular interactions at a distance of even below 0.5 nm [248]. This suggests that the two interacting proteins might be in a larger protein complex in which there are other interconnecting proteins in between.

Prior to evaluating the impact of AGR2 on EpCAM expression in cells, we first determined whether endogenous, authentically expressed AGR2 and EpCAM can form protein-protein interactions *in situ* using PLA. First, we required the use of cancer cell lines where AGR2 and EpCAM show mutual expression and where their pathways are presumably intact. In screening for cell lines that contain AGR2 and EpCAM (Figure 4.1A), we focused on the use of MCF7 cells since they express both proteins and have a wt-p53 pathway allowing for future impacts of AGR2 on p53 activity [249].

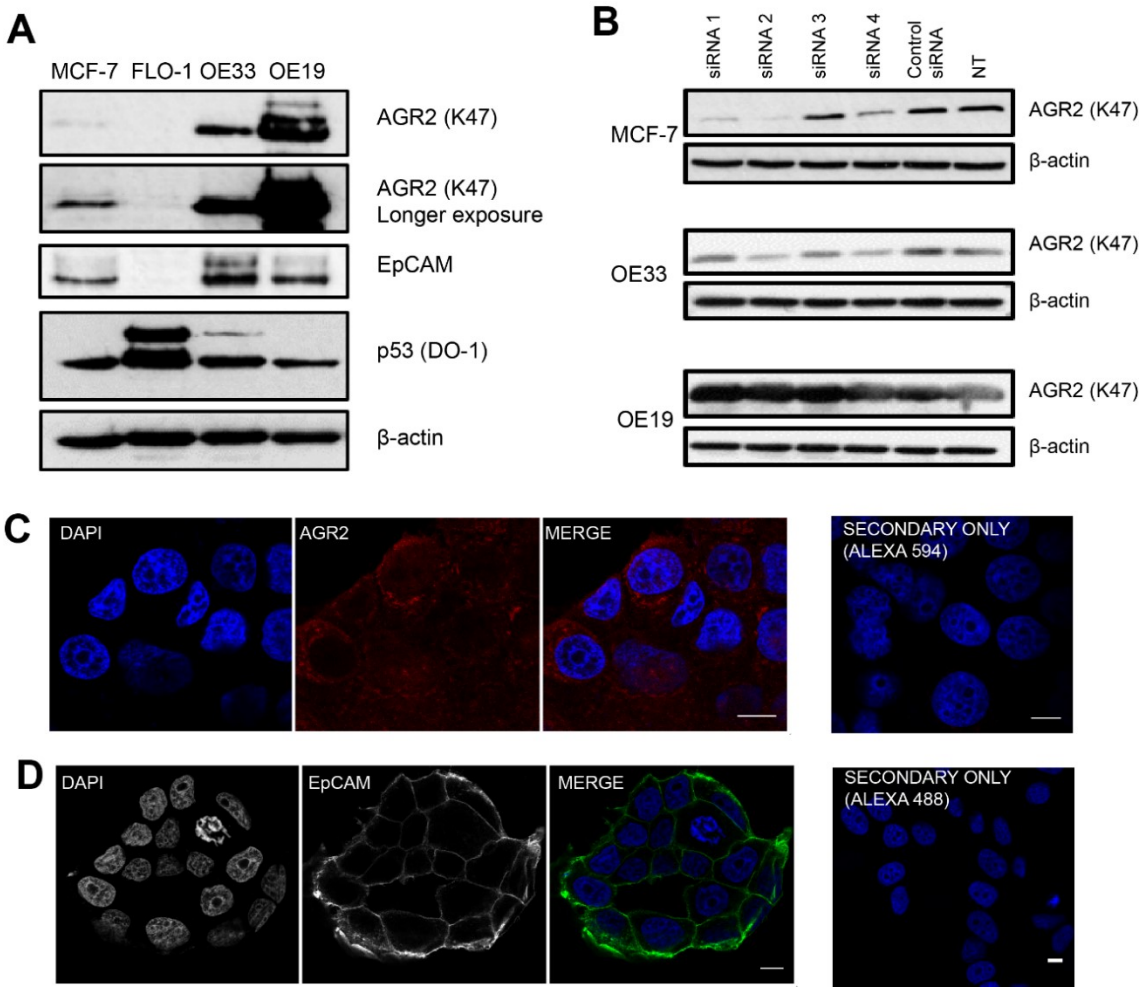


Figure 4.1 Expression of pro-oncogenic AGR2 and its putative binding partner EpCAM in a panel of cancer cell line. (A) Breast cancer cell line (MCF-7) and oesophageal cancer cell line (FLO-1, OE33 and OE19) was analysed by western blot using AGR2 polyclonal antibody K47 (Moravian Biotechnology) and EpCAM monoclonal antibody (Calbiochem). Tumour suppressor protein p53 status was also analysed p53 monoclonal antibody (Moravian Biotechnology). β-actin was used as loading control. B) Indicated cell lines were treated with four different siRNAs to AGR2 (50 µm) incubated for 72h and knockdown was detected by western blot using AGR2 polyclonal antibody K47. Control siRNA or non-transfected cells served as negative control (C) MCF-7 cells were immunostained with AGR2 polyclonal antibody K47 with a combination of Alexa 594 (red) secondary antibodies and detected using confocal microscopy. Scale bar 10 µm. (D) MCF-7 cells were immunostained with EpCAM monoclonal antibody with a combination of Alexa 488 (green) secondary antibodies and detected using confocal microscopy. Scale bar 10 µm.

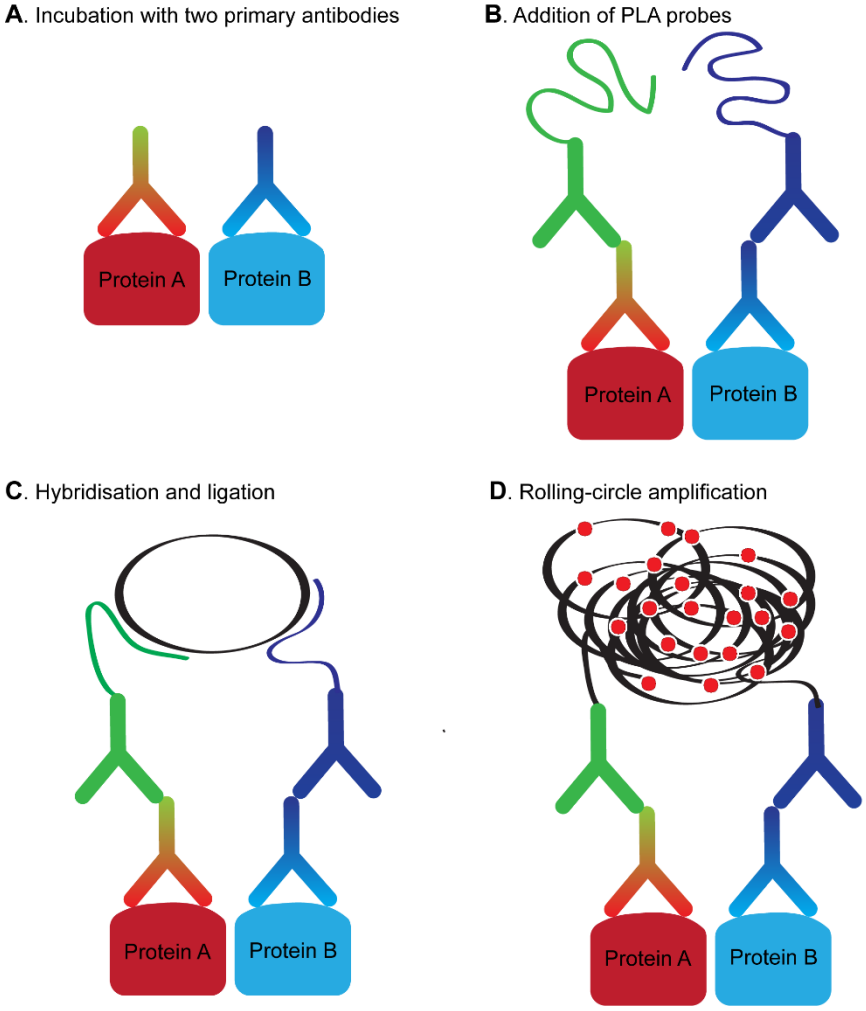


Figure 4.2 *In situ* proximity ligation assay. (A)Two primary antibodies raised in different species- for example, red protein is mouse monoclonal antibody and blue is rabbit polyclonal antibody- use to detect target antigen of interest. (B) Different short DNA strand is attached to each species-specific secondary antibody (Duolink kit) (green and blue, respectively). (C)When the secondary antibodies are in close proximity (30-40nm), the DNA strands can interact through a subsequent addition of two other circle-forming DNA oligonucleotides (black). (D) After joining of the two added oligonucleotides by enzymatic ligation, they are amplified via rolling-circle amplification using a polymerase. Amplification reaction results in several hundred replications of the DNA circle being made and labelled fluorescent probes (added as detection solution) highlight the product (red). The resulting high concentration of fluorescence in each single-molecule amplification product is easily visible as a distinct spot in a fluorescence microscope.

In addition, AGR2 has an apparently elevated dynamic in MCF7 cells since siRNA targeting AGR2 can deplete its steady-state levels (Figure 4.1B), whereas AGR2 in the p53 mutated cell lines OE19 and OE33 is more stable and perhaps less manipulated by synthetic manipulations. MCF-7 also was previously used in our lab for AGR2 studies and in fact was widely used worldwide for research purposes [250].

PLA was performed in MCF7 cells (AGR2+/EpCAM+) where the cells were incubated with antibodies to the two proteins (Figure4.3A). We tested two sets of AGR2 and EpCAM antibodies. The first pair of antibody was the antibody previously validated above in western blot and immunofluorescence which was AGR2 K47 rabbit polyclonal and EpCAM mouse monoclonal (Figure 4.3A, upper panel). The second set of antibody was AGR3 4.1 mouse monoclonal antibody (Moravian biotechnology) which showed cross-reactivity with AGR2 (Argyro Fourtouna PhD Thesis 2009) and EpCAM rabbit polyclonal (Sigma). Incubation with either pair of antibodies showed punctate foci representative of protein-protein interactions as detected by green fluorescent probes (Figure 4.3, middle panel). As a control, MCF-7 was not incubated with the antibody pair but incubated with PLA probes which showed no or few foci (Figure 4.3A, lower panel). PLA spots per cell were counted as shown in Figure 4.3C. PLA was also performed in FLO-1 cells that do not express AGR2 and EpCAM (Figure 4.3B). When authentic non-transfected FLO-1 were incubated with AGR2 and EpCAM antibodies demonstrated no significant amount of foci (Figure 4.3B and C). There are no detectable AGR2-EpCAM foci at the plasma membrane in both cell MCF-7 and FLO-1, suggesting that AGR2 interactions are confined to prior events in the maturation of the EpCAM receptor. These data together highlight that AGR2 and EpCAM can be co-expressed in cell lines (Figure 4.1 and Figure 4.2) and that a proportion of the proteins can interact in cells (Figure 4.3). These data are consistent with the prior biochemical study that mapped the *in-vitro* protein-protein contacts between AGR2 and EpCAM in Chapter 2.

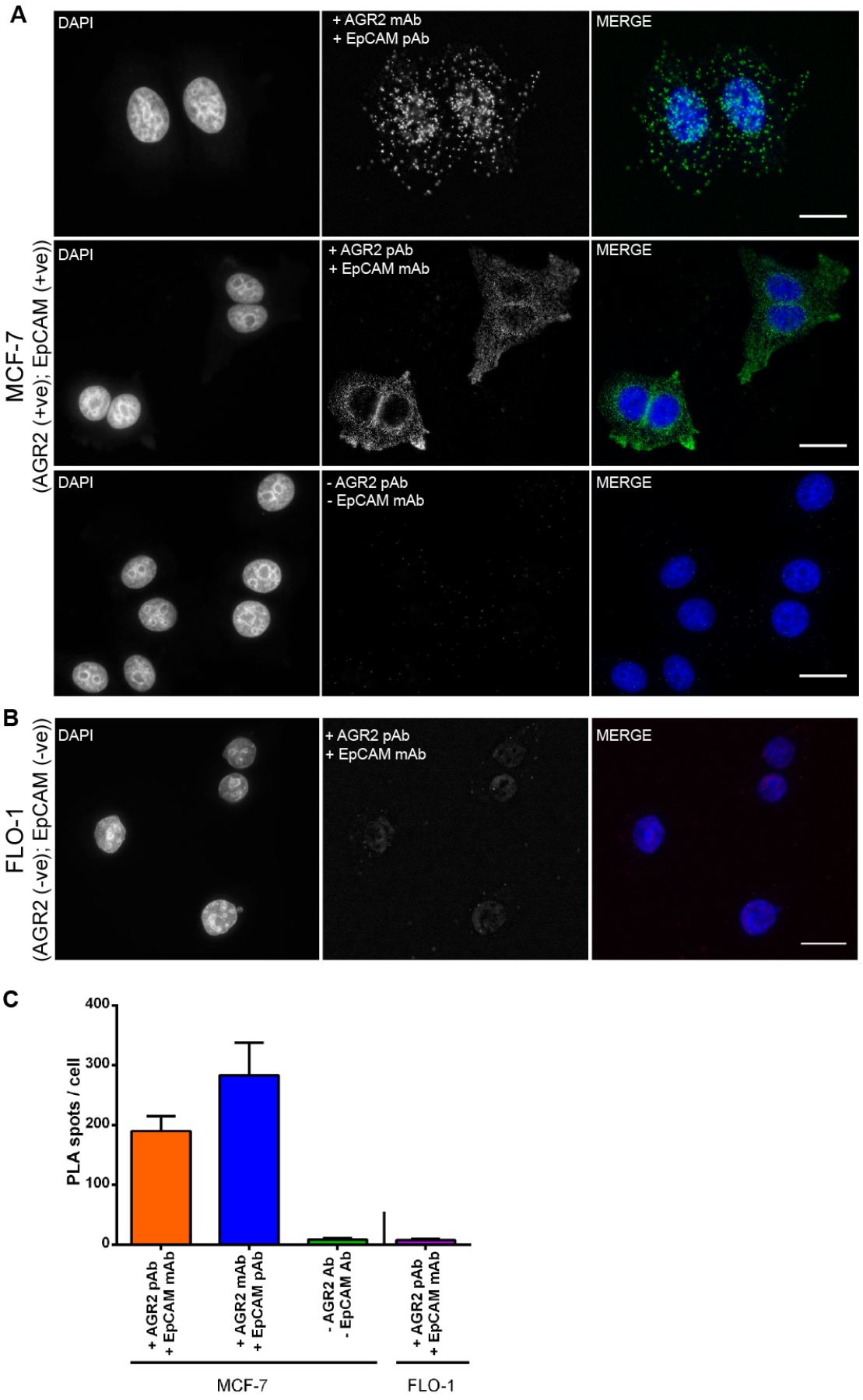

A
DAPI
+ AGR2 mAb
+ EpCAM pAb
MERGE
DAPI
+ AGR2 pAb
+ EpCAM mAb
MERGE
MCF-7
(AGR2 (+ve); EpCAM (+ve))
DAPI
- AGR2 pAb
- EpCAM mAb
MERGE
B
DAPI
+ AGR2 pAb
+ EpCAM mAb
MERGE
FLO-1
(AGR2 (-ve); EpCAM (-ve))
C
PLA spots / cell
400
300
200
100
0
+ AGR2 pAb
+ EpCAM mAb
+ AGR2 mAb
+ EpCAM pAb
- AGR2 Ab
- EpCAM Ab
+ AGR2 pAb
+ EpCAM mAb
MCF-7
FLO-1

Figure 4.3 Proximity ligation assays demonstrate that AGR2 and EpCAM form protein-protein interactions *in situ*. (A) Representative image of a proximity ligation assay performed with antibody pair of AGR2 mAB and EpCAM pAb (upper panel) or AGR2 pAb and EpCAM mAb (middle panel) in MCF-7 cells. PLA probes (Anti-rabbit PLUS probe and anti-mouse MINUS probe) were then added to the samples. Following ligation and amplification, protein-protein interaction complex was detected with green fluorescent probes (Duolink). Green fluorescence foci indicate the interaction between the two proteins. As a control, MCF-7 was not incubated with the antibody pair but incubated with PLA probes (lower panel) which showed no or few foci. Scale bar 25 µm (B) PLA also was also performed in cells that do not express AGR2 and EpCAM (FLO-1). AGR2 pAb and EpCAM mAb was used and protein-protein interaction complex was detected with red fluorescent probes (Duolink). Transfected AGR2 and EpCAM in FLO-1 cells showed foci indicative of protein-protein interaction while authentically non-transfected FLO-1 demonstrated no significant amount of foci. Scale bar 10 µm. Nuclei were counterstained with DAPI and cells were visualised with an epifluorescence microscope. (C) Signals were quantified in randomly selected 20 cells using ImageJ software (analyse particles) and error bars show SEM. The experiment was repeated three times and a representative experiment is shown.

4.2.3. Co-localisation analysis of AGR2 and EpCAM by fluorescence microscopy

We next aim to see if AGR2 and EpCAM can be seen co-localise in cells. Co-localisation is defined as the presence of two or more different molecules residing at the same physical location in a cell. If proteins interact *in-vivo* they are expected to be co-localised or, at least, they will display overlapping distribution within the cell. The intracellular localisation of two (or more) proteins can be studied by fluorescence microscopy. In the classical immunofluorescence, cells are incubated with two primary specific antibodies, followed by incubation with secondary antibodies conjugated to two different fluorophores. Co-localisation is a function of two fluorophores display different emission maxima, the intracellular localisation of the proteins can be monitored by fluorescence microscopy. Therefore, if the two proteins are co-localised the fluorescent probes will also be co-localised which are represented by pixels from fluorescence microscope containing both colour distributions.

The most commonly used method to comprehend co-localisation is immunofluorescence (IF) where specific antibodies for protein and its binding partner were incubated in cells followed by secondary antibody conjugated with two different fluorophores. However, there are two important problems related to this method. First, many antibodies can bind non-specifically to other proteins or show cross-reactivity with other proteins. Second, the method requires antigen retrieval which involves the fixation and permeabilization of the cells that can potentially cause artefacts that can affect the localisation observed. Many studies have conveyed different results in protein distribution depending on the antigen retrieval protocol used [251]. Since *in-situ* localisation of proteins is a function of maintaining '*in-vivo'* protein distribution after immunostaining, protein redistribution needs to be avoided and extraction of the antigen during experimental needs to be augmented.

Although we have shown the specificity of the AGR2 and EpCAM using western blot and immunofluorescence studies (Figure 4.1C and D), the use of fusion protein will ensure minimum artefacts during immunolabelling and eliminates the use of antigen retrieval protocol which could necessitate protein distribution. Full-length

AGR2 (FL-AGR2) was previously cloned into mCherry vector and a gift from Dr Veronika Brychtova (Brno, Czech Republic) (Figure 4.4A). mCherry is a mutant fluorescent protein derived from the tetrameric *Discosoma sp.* red fluorescent protein, DsRed. Full-length EpCAM (FL-EpCAM) was cloned into a pEGFP-N1 vector which encodes a red-shifted variant of wild-type GFP which has been optimised for enhanced fluorescence and higher expression in mammalian cells (Figure 4.4A). pEGFP-N1 encodes the GFPmut1 variant which contains the double-amino-acid substitution of Phe-64 to Leu and Ser-65 to Thr. Both FL-AGR2 and FL-EpCAM harbour C-terminus tagged of mCHERRY and EGFP respectively. From herein, we use AGR2-CHERRY and EpCAM-GFP for describing these constructs.

The expression of the fusion proteins was then observed in cells using western blot to make sure they expressed properly. AGR2-CHERRY or EpCAM-GFP DNA constructs were transfected into MCF-7 cells. As a control, MCF-7 cells also were transfected with mCHERRY and GFP-expressing vectors and were immunoblotted using an mCHERRY antibody or GFP antibody respectively. Data showed that full-length mCHERRY-AGR2 was expressed (arrow). However, we noticed the reproducible small molecular mass “cleavage” or synthesis product when mCHERRY was transfected into cells (asterisk). The immunoblotting of GFP and EpCAM-GFP transfected cells using GFP antibody highlights the expression of GFP alone marked by an asterisk (lane 1) and full-length EpCAM-GFP fusion protein marked by an arrow (lane 2). Interestingly, The EGFP was not subjected to the production of smaller molecular mass adducts as was the mCHERRY protein, which suggests that GFP fluorescent tags are more stable in cells.

Using confocal microscopy, transfection of AGR2-CHERRY into MCF-7 showed cytosolic expression. This localisation of the tagged AGR2 was consistent with IF data, but with tagged-AGR2 the punctate are more vivid and clear (Figure 4.5C). Transfected EpCAM-GFP in MCF-7 showed major localisation in the membrane consistent with the nature of the protein being a transmembrane protein (Figure 4.5D) and consistent with IF results. The mCHERRY controls and GFP controls are shown in Figure 4.5 A and B, respectively which showed non-specific

localisation. When both DNA constructs were co-transfected in MCF-7 cells, a variable degree of colocalisation was observed (Figure 4.6 A). Two representative cells were pooled to showed the different co-localizations. AGR2 and EpCAM showed partial co-localisation in the cytoplasmic region (Figure 4.6A). By contrast, the EpCAM^{Y251A}-GFP mutant with an attenuated AGR2 binding previously shown using *in-vitro* biochemical study in Chapter 3 expressed cytosolically (Figure 4.5E). Co-transfection of the mutant EpCAM and AGR2-CHERRY showed that co-localisation is still retained but to a lesser extent (Figure 4.6B). However, there is also the observation that there is more co-localisation of the mutant EpCAM and AGR2 (Figure 4.6B, upper panel). We think that maybe this is due to mutant EpCAM is unfolded and therefore accumulating in the ER where AGR2 predominantly resides.

Altogether, these data suggest that the transfection of AGR2 and EpCAM can result in transient interactions and is consistent with the pool of endogenous interactions identified using proximity ligation (Figure 4.3). Co-localisation of AGR2 and EpCAM in the cytoplasmic region suggest that the interaction event could either be in the ER, the first compartment of the secretory pathway, where AGR2 chaperone EpCAM for proper folding or co-localisation events occur during trafficking (e.g Golgi apparatus, endosome constituents) for eventual EpCAM membrane delivery. Using this cell line with a dynamic AGR2 protein expression, we first noted using confocal microscopy that the majority of AGR2 is cytosolic and EpCAM is plasma membrane (Figure 4.5C and D), which is the expected dominant localisation of the proteins. By contrast, the EpCAM^{Y251A} mutant with attenuated AGR2 binding *in-vitro* is distributed away from the plasma membrane but still reveals co-localisation with AGR2 (Figure 4.5E and Figure 4.6B). When the mutant EpCAM was tested on binding to AGR2 *in-vitro*, the binding was not fully abolished, which explained why there is still evidence of co-localisation observed in cells.

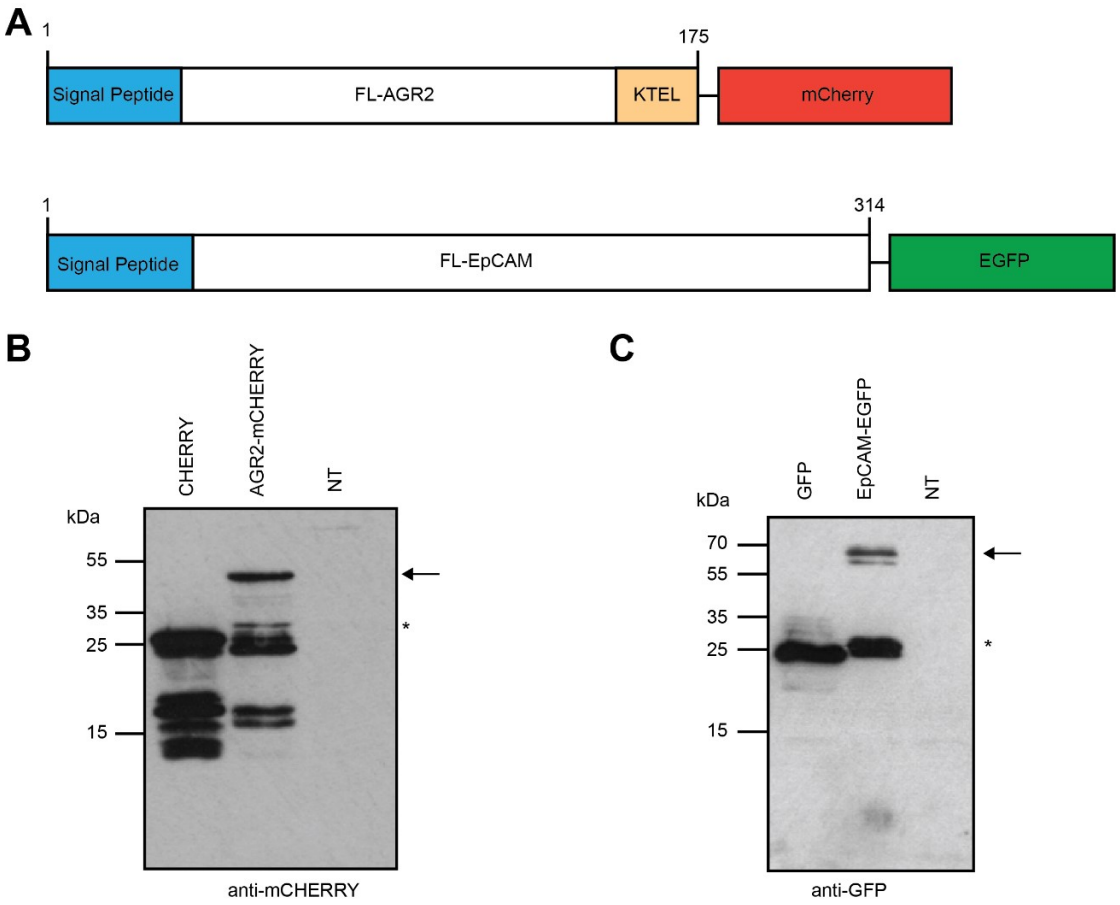


Figure 4.4 Constructions of AGR2-CHERRY and EpCAM-GFP fusion proteins. (A) Schematic representation showing constructs of fluorescently labelled versions of AGR2-CHERRY and EpCAM-GFP. Both proteins contain the full-length sequences. (B) The immunoblotting of mCHERRY and mCHERRY-AGR2 transfected cells highlights the expression of mCHERRY alone (lane 1) and mCHERRY-AGR2 (lane 2). Blots were incubated with an anti-mCHERRY antibody. The arrow marks the location of full-length mCHERRY-AGR2 and the asterisk marks the location of mCHERRY. We noticed the reproducible small molecular mass "cleavage" or synthesis products when mCHERRY was transfected into cells. (C) The immunoblotting of EGFP and EpCAM-GFP transfected cells highlights the expression of EGFP alone (lane 1) and EpCAM-GFP (lane 2). Blots were incubated with a GFP antibody. The arrow marks the location of full-length EpCAM-GFPand the asterisk marks the location of GFP. The GFP was not subjected to the production of smaller molecular mass adducts as was the mCHERRY protein. NT= Non-transfected.

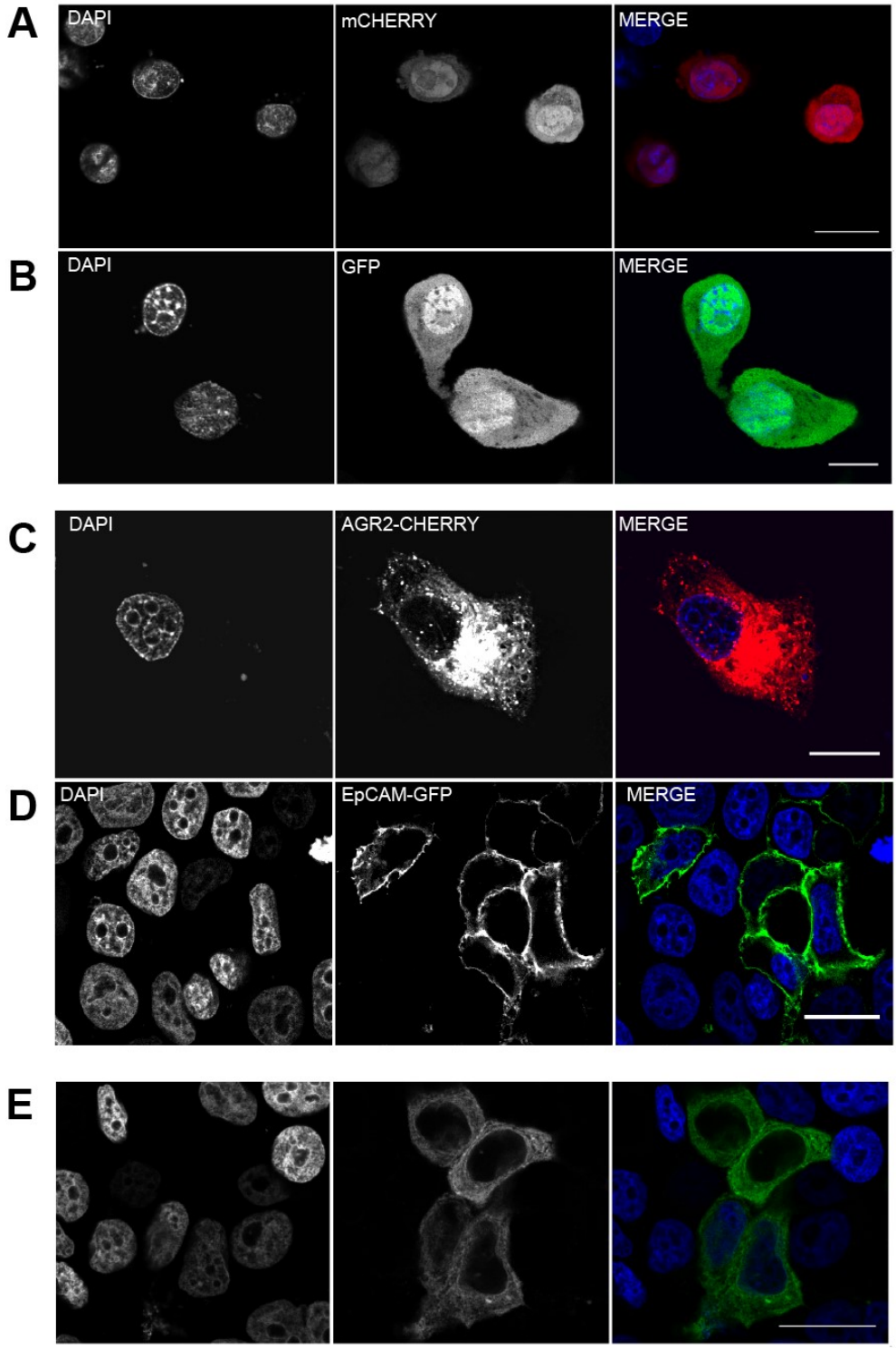


Figure 4.5 Expression and localisation analysis of AGR2-CHERRY and EpCAM-GFP in MCF-7 cells using confocal microscopy. A) MCF-7 cells transfected with mCHERRY vector or GFP vector demonstrate non-specific expression. C) MCF-7 cells transfected with AGR2-CHERRY. AGR2-CHERRY localised in the perinuclear region and cytoplasm as punctate. D) MCF-7 cells transfected with EpCAM-GFP showing partial cytoplasmic localisation and but mainly localise in the cell membrane. E) Effect of AGR2 peptide docking site loss-of-function mutation. EpCAM-GFP were mutated at Y251A. Transfected EpCAM^{Y251A}-GFP de-localize to the cytosol. All cells were analysed 24-hours after transfection and cells were fixed using 4% PFA. Nuclei were counterstained with DAPI. Scale bar 25 µm.

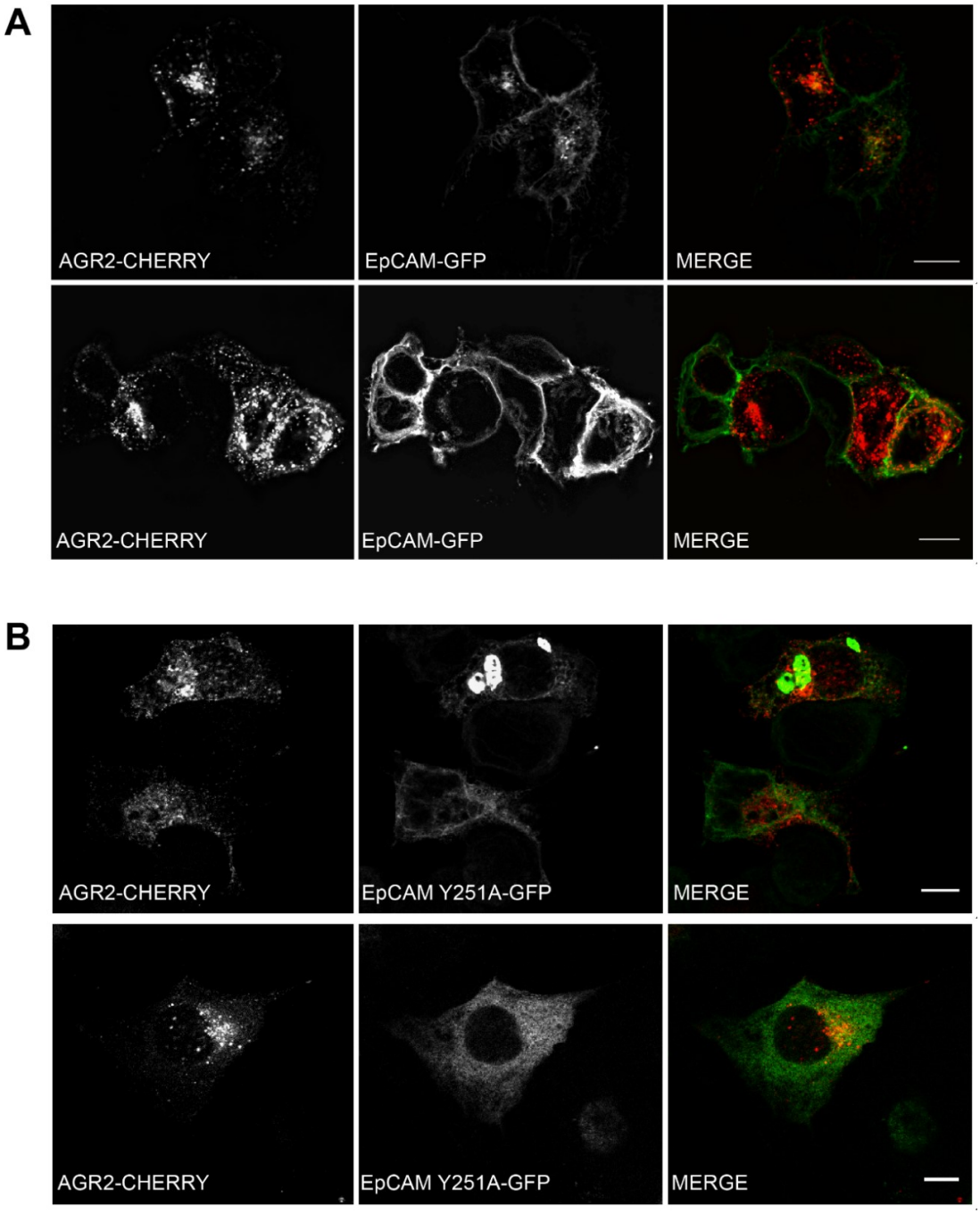


Figure 4.6 Co-localisation of AGR2-CHERRY and EpCAM-GFP using confocal microscopy. (A) MCF-7 cells were transfected with AGR2-CHERRY and EpCAM-GFP and visualised using confocal microscopy. Two representative images were shown to show a degree of co-localisations. Scale bar 10um. (B) MCF-7 cells were transfected with AGR2-CHERRY and EpCAM^{Y251A}-GFP and visualised using confocal microscopy. Two representative images were shown to show a degree of co-localisations. Scale bar 10 µm. All cells were analysed 24-hours after transfection and cells were fixed using 4% PFA.

4.2.4. Reconstitution of an AGR2-EpCAM pathway dynamics using cell-based assays.

Next, we aimed to exploit FLO-1 cells, which has no detectable AGR2 or EpCAM expression (Figure 4.1A), as a potential assay that can reconstitute AGR2 impacts on EpCAM expression. We ask whether AGR2 docking to EpCAM has a positive or negative role in receptor maturation, expression, or localisation. The mislocalisation of the EpCAM-^{Y251A} mutant suggests that attenuated binding results in this mislocalisation of EpCAM from going to the plasma membrane. So, we might expect to develop an assay whereby AGR2 transfection can stimulate EpCAM membrane localisation. The same approach was used recently to study AGR2 and its binding partner, epidermal growth factor receptor (EGFR) [111]. Like EpCAM, EGFR is a membrane protein. This served another good example of how an ER-based thioredoxin AGR2 is able to direct signalling activity by influencing whether EGFR is delivered to the plasma membrane from the secretory pathway. Dong and colleagues used AGR2 and EGFR negative cells (CHO-K1) to show AGR2-induced EGFR delivery to the plasma membrane [111]. Transient transfection of CHO-K1 cells with an EGFR-GFP construct alone resulted in cytosolic distribution. Co-transfection with EGFR-GFP and AGR2 cDNA resulted in EGFR cell surface expression. So, we ask whether this is the case with AGR2 and EpCAM after having found cell model that is negative for both proteins.

Initially, we sought to transfect EpCAM-GFP alone with two different concentrations (low and high amount of DNA) to see if there any difference in localisation compare to EpCAM positive cells, MCF-7 (Figure 4.7). Initial transfection of low amount EpCAM-GFP in FLO-1 cells showed that the majority of the localisation is cytosolic (Figure 4.7, middle panel). Transfection of a higher amount of EpCAM-GFP DNA constructs showed almost similar results where most EpCAM-GFP are cytosolic but there are some mixtures of cytosolic and diffused membrane staining (Figure 4.7, lower panel). There were no differences when AGR2 were transfected with the low and high amount of AGR2-CHERRY, where they showed typical cytosolic distribution near perinuclear region which was similarly seen in MCF-7 cells (Figure 4.7B, middle and lower panel). The empty vector encoding EGFP

control of mCHERRY were shown in (Figure 4.7A and B, upper panel) which showed non-specific distribution. Co-transfection of an equal amount of AGR2-CHERRY and EpCAM-GFP DNA into FLO-1 cells resulted in more membrane expression of the EpCAM-GFP per field views (Figure 4.8B and enlarged image) compared to EpCAM-GFP alone (Figure 4.8A). These initial data suggested that AGR2 is required for EpCAM delivery to the plasma membrane. Next, we ask whether we can optimise the assay to elevate EpCAM delivery to the plasma membrane by finding the best ratio of AGR2 to EpCAM DNAs transfected into FLO-1 cells. We performed several titrations of AGR2-CHERRY and EpCAM-GFP co-transfection, where EpCAM-GFP was kept at a lower concentration with increasing amount AGR2-mCHERRY to observe whether AGR2 has an impact on trafficking EpCAM to the plasma membrane. As the ratio of AGR2-CHERRY to EpCAM-GFP increased, EpCAM membrane staining was more evident(Figure 4.9A). We counted the cells with either; i) cytoplasmic positive and membrane negative or ii) membrane positive in a field containing at least 50 cells and the counts were plotted in Figure 4.9B. The graph shows a shift (%) from cytoplasmic staining to membrane staining as AGR2-CHERRY increases(Figure 4.9B).The best ratio of AGR2-mCHERRY to EpCAM-GFP, where EpCAM-GFP was most seen in the plasma membrane, is at 3:1 (or 1.5 µg: 0.5 µg per 6 well plates). When the experiment was repeated using the ratio, EpCAM-GFP can be seen more evident in membrane compare to cytoplasmic localisation (Figure 4.9C), and also suggest reproducibility of EpCAM delivery to the membrane upon co-transfection with AGR2.

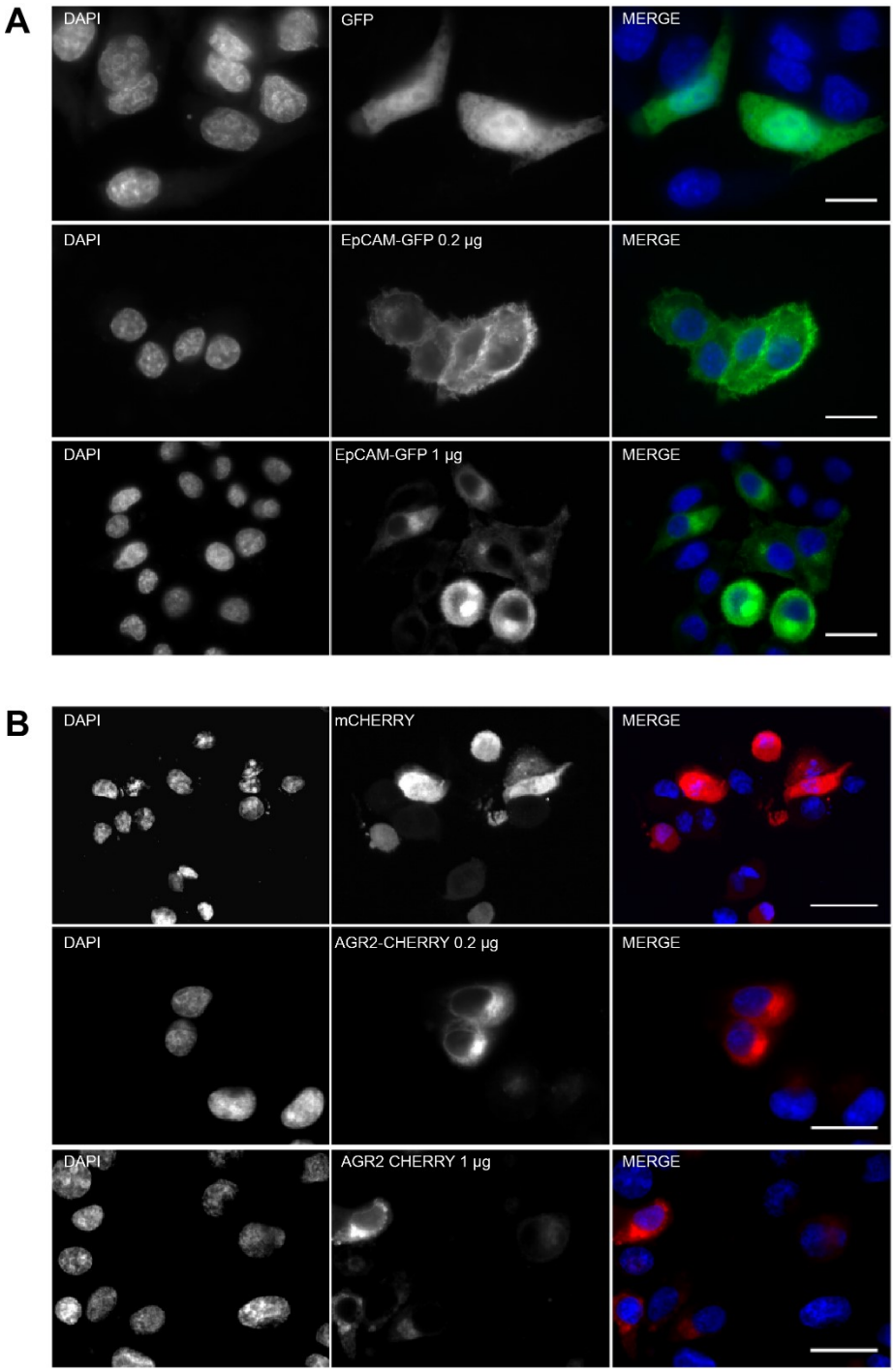


Figure 4.7 Localisation of AGR2-CHERRY and EpCAM-GFP in FLO-1 cells (AGR2 (-ve); EpCAM -ve). (A) FLO-1 cells were transfected with GFP vector only (upper panel) low amount of EpCAM-GFP, 0.2µg (middle panel) and high amount of EpCAM-GFP 1 µg (lower panel). Scale bar 25 µm. (B) FLO-1 cells were transfected with mCHERRY vector only (upper panel) low amount of AGR2-CHERRY, 0.2µg (middle panel) and high amount of AGR2-CHERRY 1 µg (lower panel). Scale bar 25 µm. Nuclei were counterstained with DAPI and cells were visualised with an epifluorescence microscope.

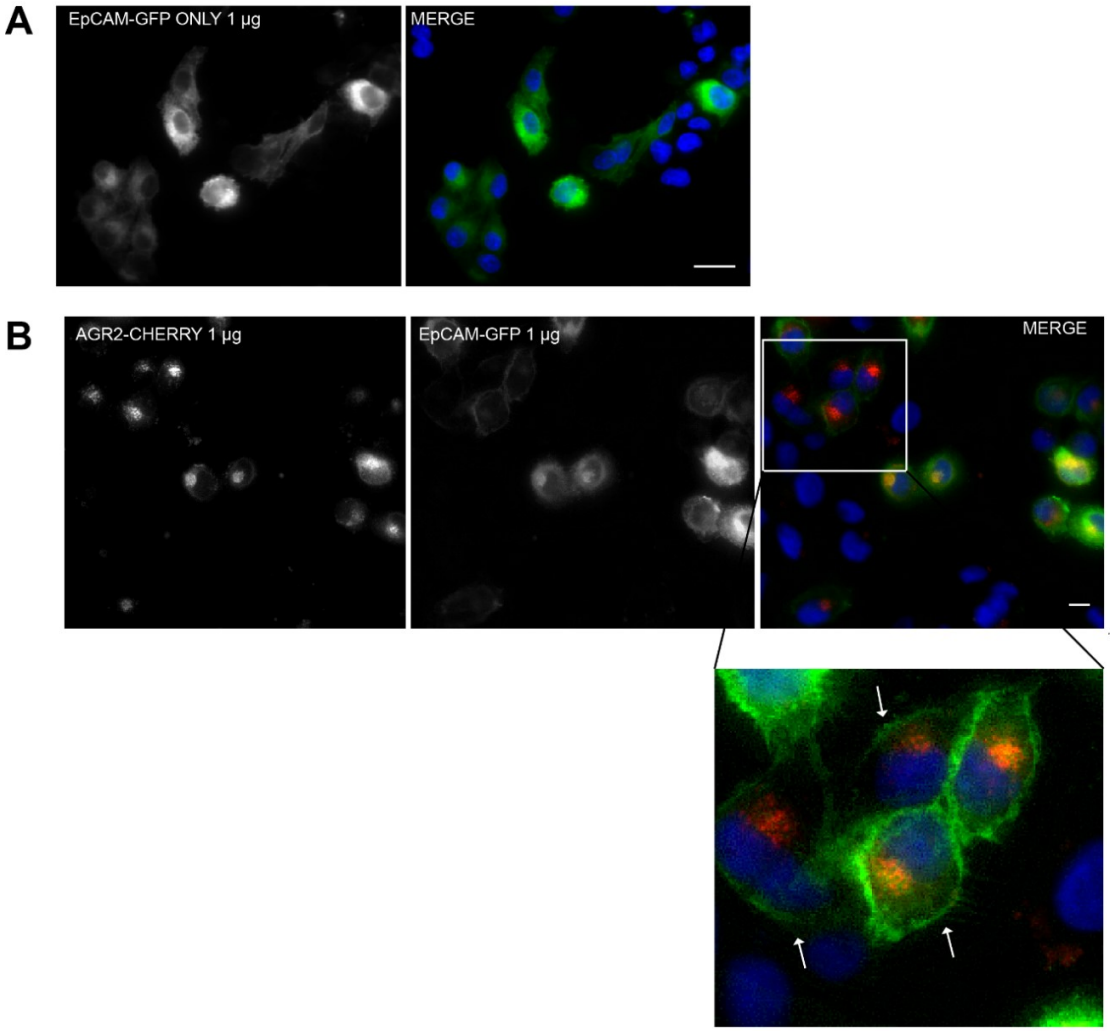
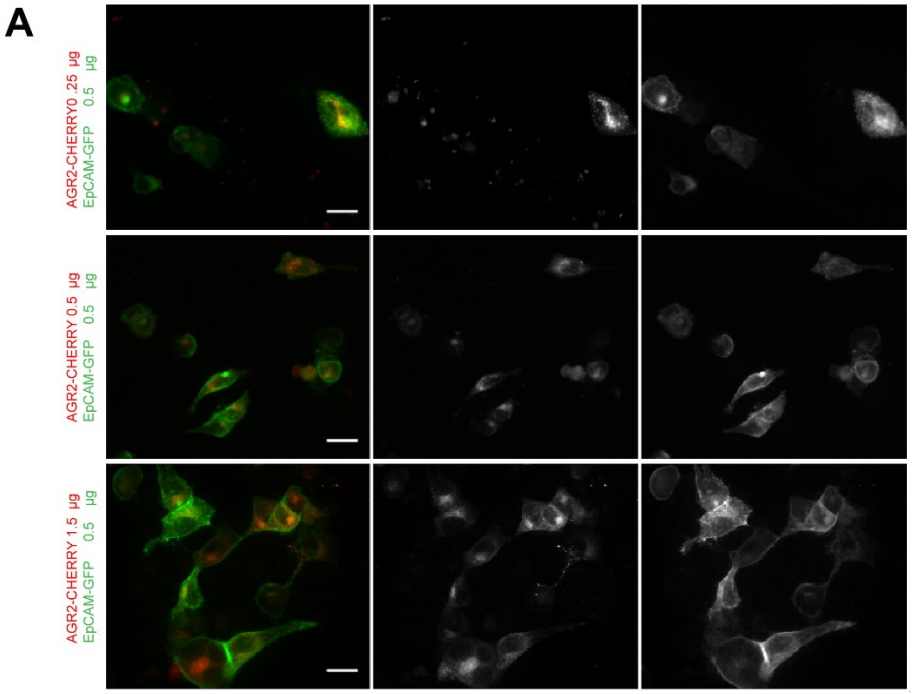
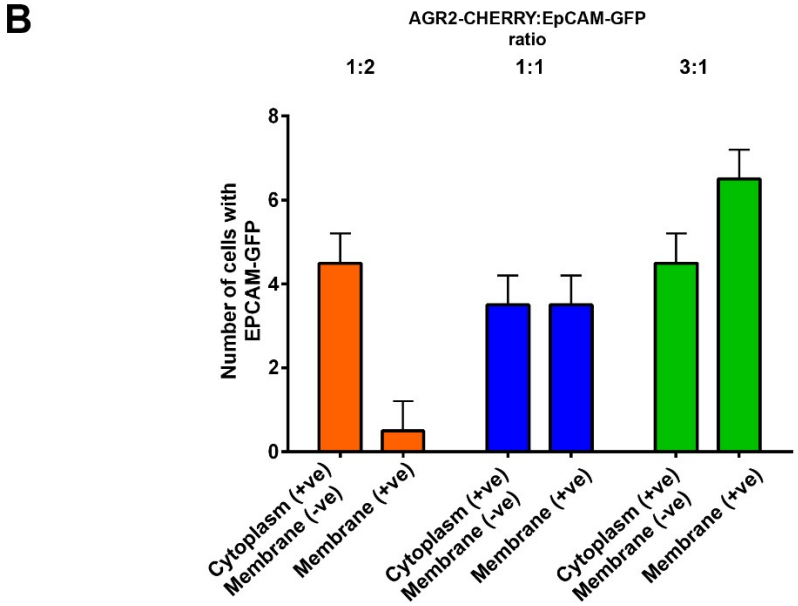


Figure 4.8 Initial co-transfection of AGR2-CHERRY and EpCAM-GFP into FLO-1 cells. (A) FLO-1 were transfected with EpCAM-GFP only. Scale bar 10 µm (B) FLO-1 cells were co-transfected with an equal amount of AGR2-CHERRY and EpCAM-GFP DNA. EpCAM-GFP with more membrane localisation in the presence of AGR2-mCHERYY was shown in the enlarged image. Scale bar 10 µm. Nuclei were counterstained with DAPI and cells were visualised with an epifluorescence microscope.


A
AGR2-CHERRY0.25 µg
EpCAM-GFP 0.5 µg
AGR2-CHERRY 0.5 µg
EpCAM-GFP 0.5 µg
AGR2-CHERRY 1.5 µg
EpCAM-GFP 0.5 µg


B
AGR2-CHERRY:EpCAM-GFP ratio
1:2
1:1
3:1
Number of cells with EPCAM-GFP
0
2
4
6
8
Cytoplasm (+ve) Membrane (-ve)
Membrane (+ve)
Cytoplasm (+ve) Membrane (-ve)
Membrane (+ve)
Cytoplasm (+ve) Membrane (-ve)
Membrane (+ve)

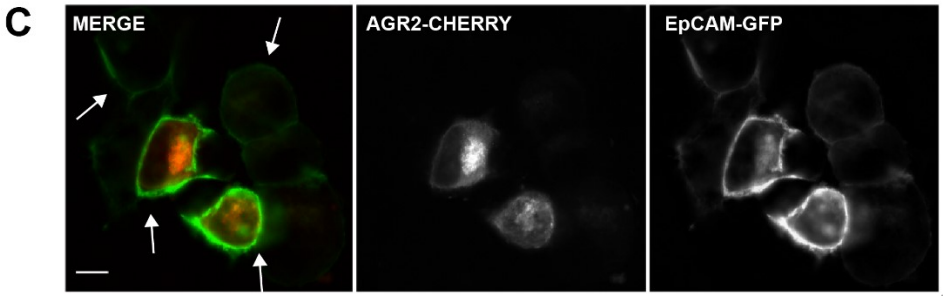

C
MERGE
AGR2-CHERRY
EpCAM-GFP

Figure 4.9 EpCAM-GFP enhance membrane localisation with titration of AGR2-CHERRY in FLO-1 cells. (A) FLO-1 cells were co-transfected with a fixed amount of EpCAM-GFP (0.5 µg) and AGR2-CHERRY titration (0.5-1.5 µg). Scale bar 10 µm. (B) Graph showing the number of cells with either cytoplasmic positive/membrane negative or membrane positive with each titration as in (A). Widefield view of at least 10 cells per field was counted. (C) FLO-1 cells were -transfected in an independent experiment with AGR2-CHERRY (0.5 µg) and EpCAM-GFP (1.5 µg) or 3:1 ratio which showed enhanced membrane staining.

In addition, upon quantitation of the extent of EpCAM-GFP membrane vs cytosolic localisation, we noticed elevated fields of EpCAM-GFP expression in wide-field views. To confirm this, we co-transfected very low amount of EpCAM-GFP (0.125 µg DNA) with titration of AGR2-CHERRY (0.25 µg- 1.5 ug DNA). The increasing amount of AGR2-CHERRY DNA elevates the production of lowly transfected EpCAM, while the transfection of the low EpCAM-GFP alone was poorly expressed (Figure 4.10A). Quantitation of cells with EpCAM-GFP positive cells per field views confirm elevated EpCAM-GFP when co-transfected with increasing AGR2-CHERRY expressed (Figure 4.10B). This suggests that AGR2 can elevate the steady-state levels of EpCAM. This EpCAM elevation was confirmed using immunoblotting (Figure 4.11). When the same experiment was repeated and the lysates were subjected to immunoblotting, the bands corresponding to EpCAM increased as AGR2-CHERRY increases (Figure 4.11A, upper panel) and confirmed when the integrated intensity of bands was quantified (Figure 4.11A, lower panel). We also evaluated whether the sole cysteine (Cys-81) within the thioredoxin domain on AGR2 has any role in AGR2: EpCAM interaction. The cysteine residue was previously shown to form mixed disulphide bridge with AGR2 client proteins such as MUC2 and EGFR [60, 111]. We mutated AGR2-CHERRY at its sole cysteine within the to Cys81 to Ser (AGR2-^{C81S}-CHERRY). However, there was no significant difference from WT-AGR2 in stimulating EpCAM production suggesting that a chaperonin function rather than a redox function is directly responsible for EpCAM elevation (Figure 4.11).

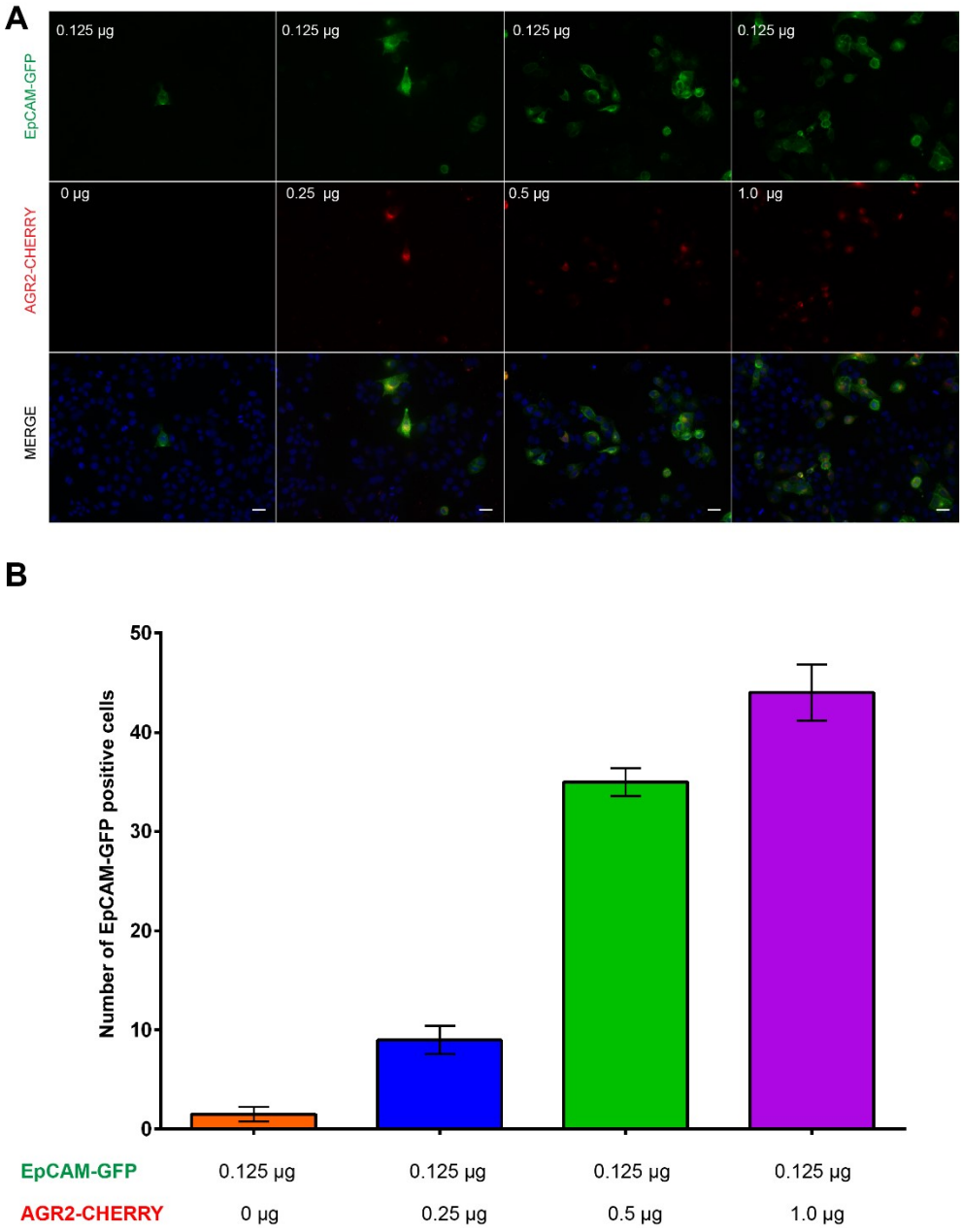


Figure 4.10 EpCAM-GFP expression in FLO-1 cells were elevated with increasing AGR2-CHERRY. (A) FLO-1 cells were either transfected with relatively low amount of EpCAM-GFP alone or co-transfected with a fixed amount of EpCAM (0.125 µg) and titration of AGR2-CHERRY (0.25 µg – 1 µg). Transfection of EpCAM-GFP alone showed low expression and barely detectable. Upon titration of AGR2 co-transfected with EpCAM, the expression of EpCAM is greatly increased. Scale bar 25 µm. Nuclei were counterstained with DAPI and cells were visualised using epifluorescence microscope.

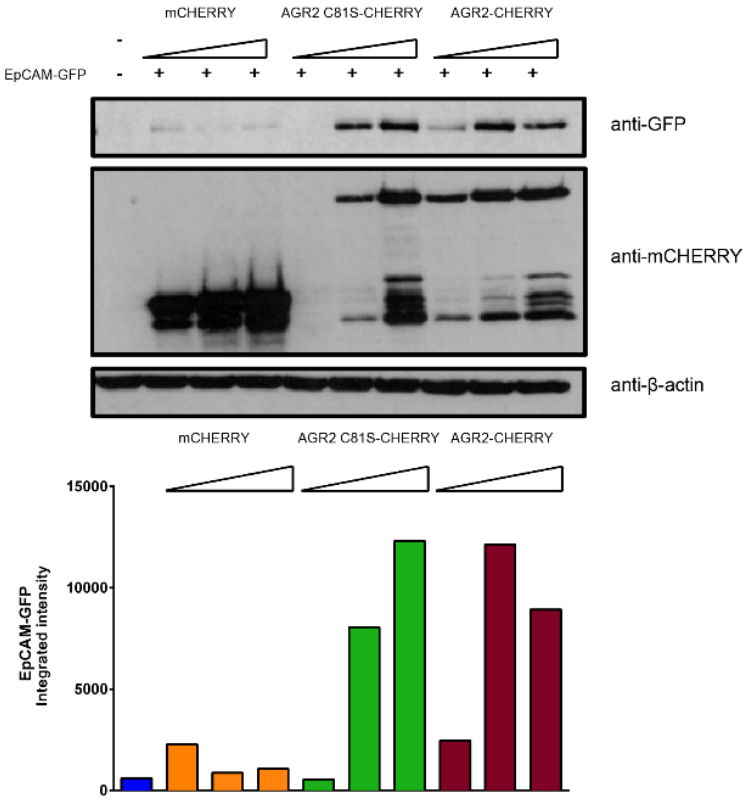


Figure 4.11 Western blot showing EpCAM-GFP expression in FLO-1 cells were elevated with titration of AGR2-CHERRY. FLO-1 cells were either transfected with a low amount of EpCAM-GFP alone or co-transfected with a fixed amount of EpCAM (0.125 µg) and in a combination with increasing amount of mCHERRY, AGR2-CHERRY or AGR2^{C81S}-CHERRY (0.25 µg, 0.5 µg, 1 µg) and analysed using western blot with Anti-GFP antibody to detect the presence of EpCAM-GFP. Only the band correspond to full-length AGR2-GFP was shown. The anti-mCHERRY antibody was used to detect the presence of mCHERRY, AGR2-CHERRY or AGR2^{C81S}-CHERRY. β-actin was used as loading control. Integrated intensity corresponding to the full-length EpCAM-GFP band was scored as a bar graph (lower panel).

4.2.5. Construction of recombinant untagged-AGR2 and untagged-EpCAM in weak and strong promoter vectors

We have shown that AGR2 can elevate the steady-state level of EpCAM using fluorescence microscopy and western blot using fusion proteins where both proteins were fluorescently tagged. We asked whether the fluorescent tags could have an effect to the elevation of EpCAM. Hence, we thought of making the assay cleaner by removing the tags in both proteins to ensure that genuine AGR2 causing the effect of elevating EpCAM expression. We cloned AGR2 and EpCAM into two different mammalian vectors: strong and weak constitutive promoters. This represents the low and high expression of AGR2 or EpCAM and also eliminates the need of titration. The strong constitutive promoter contains human cytomegalovirus (CMV) promoter (pSF-CMV) which is by far the strongest mammalian and the most commonly used promoter (Figure 4.12B). The weak constitutive promoter was derived from Rous sarcoma virus (RSV) promoter (pSF-RSV) (Figure 4.12B). Both of the vectors were bought from Oxford Genetics which the expression levels of different mammalian promoters have been tested. Figure from Oxford Genetics[13] showed that CMV demonstrated the highest transient expression level in mammalian cells whereas RSV showed the weakest expression level (Figure 4.12A). While all the DNA sequences are similar, the two vectors were only differing in the promoter region and antibiotic selection gene where the pSF-CMV vector has ampicillin resistance gene while the pSF-RSV has kanamycin resistance gene (Figure 4.12B).

[13] www.oxfordgenetics.com

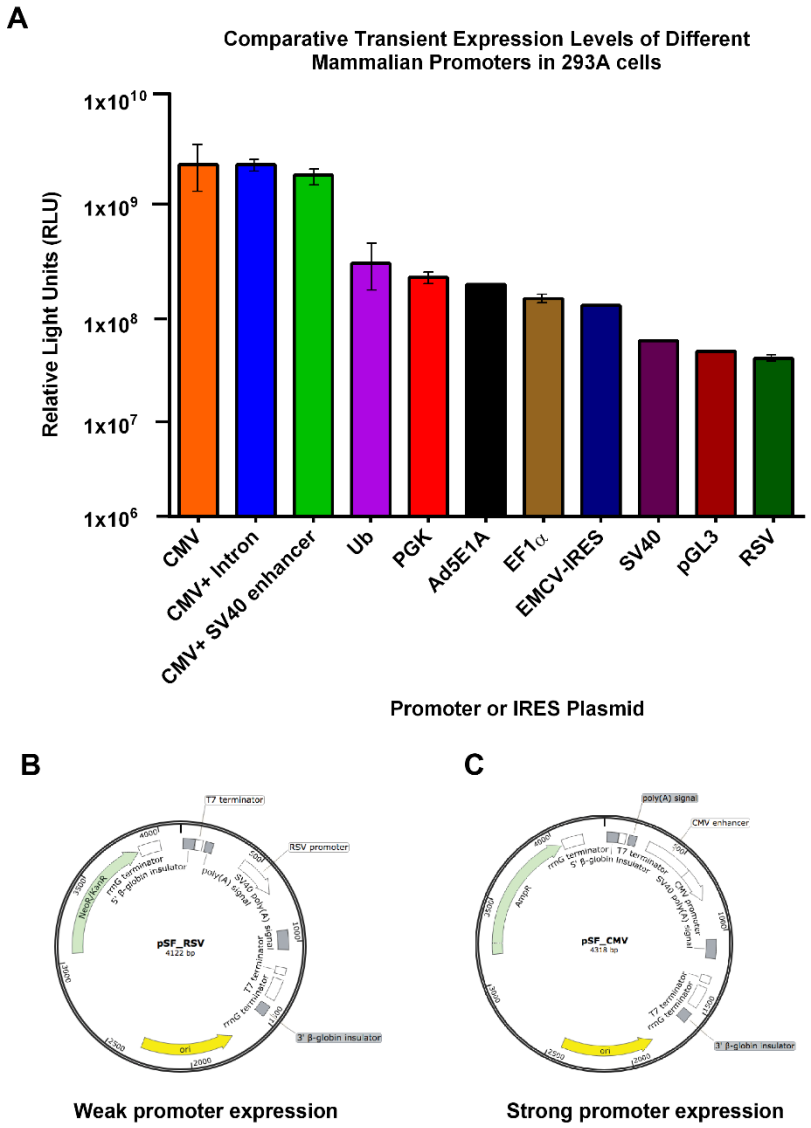


Figure 4.12 Comparison of the transient expression level of different mammalian promoters. (A) Bar Graph showing relative expression levels of different mammalian promoters tested in 293A cells (reproduced from OxfordGenetics.com). (B) pSF-RSV vector which contains Rous sarcoma virus (RSV) promoter that drives gene expression in mammalian cells with a kanamycin resistance cassette. This promoter is one of the weaker constitutive promoters as shown in (A). (C) pSF-CMV vector which contains human cytomegalovirus (CMV) promoter that drives gene expression in mammalian cells with an ampicillin resistance cassette. This promoter is one of the strongest constitutive promoters as shown in (A). Both vectors contain few important sites within the multiple cloning sites (MCS). These include the NcoI site the XbaI site and the BsgI and BseRI sites. The NcoI site contains a start codon that is immediately downstream of both a Kozak and Shine-Dalgarno ribosomal binding site which allow for optimal positioning of genes when the start codon is placed in this location. The XbaI site contains a stop codon. This stop codon is positioned in a specific position in relation to the BsgI and BseRI sites that are immediately downstream. When either BseRI or BsgI cleave the plasmid they produce a TA overhang from the stop codon in the XbaI site.

4.2.5.1. Cloning and expression of structural variants of untagged-AGR2

We also interested to see if different structural variants or molecular arrangement of AGR2 published in the literature have a positive or negative impact on EpCAM expression. Based on previous studies, full-length and mature form (signal peptide aa 1-21) of AGR2 showed different localisation in cells. Mutagenesis of ER retention site of KTEL to more optimise KDEL showed more ER distribution in cells [81]. The mature form of AGR2 with deletion of KTEL showed that AGR2 is secreted outside the cells [81, 96]. When the intrinsically disordered N-terminal region (aa 21-45) was deleted, AGR2 exists in a more stable and spontaneous dimer when in the purified form [97]. Based on this information, we decided to test four different AGR2 forms on EpCAM activity in which we designed vectors containing i) full-length AGR2 WT (FL-AGR2 WT); ii) full-length AGR2 with T173D mutation (FL-AGR2 KDEL - enhanced ER); iii) mature AGR2 + KTEL deletion (mAGR2 ΔKTEL-enhanced secretion); iv) N-terminal deletion aa 20-40 of AGR2 + KDEL (ΔNterm AGR2 KDEL - enhanced ER dimer). Schematic representation of these AGR2 forms and their description were shown in Figure 4.13A and B. The cDNAs of these were chemically synthesised along with restriction enzymes for directional cloning into low (pSF-RSV) and high expression (pSF-CMV) vectors. The 5' and 3' ends of each cDNA contain *NcoI* and *XbaI* restriction sites respectively and cloned into the multiple cloning sites containing the same restriction sites in pSF-RSV and pSF-CMV. Resulting recombinant AGR2 constructs were named as shown in Figure 4.13B.

We then checked how these different forms of AGR2 express in the cells and whether the weak and strong promoters give different level of expression. We used FLO-1 cells as this cells will be used for assaying AGR2: EpCAM interaction. FLO-1 cells transfected with FL-AGR2 WT and FL-AGR2 KDEL under weak RSV promoter and analysed in western blot using AGR2 antibody gave a single band, with the later showed a higher band intensity (Figure 4.13C). The mAGR2 ΔKTEL showed almost undetectable band while ΔNterm AGR2 KDEL demonstrated two bands (Figure 4.13C). AGR2 constructs under the control of the strong promoter, CMV showed a similar pattern of bands but with increased band intensities (Figure 4.13C). This

suggests that the two promoters indeed showed the expected level of expression, where RSV promoters drive weaker expression of AGR2 while CMV gives higher AGR2 expression.

The higher expression of FL-AGR2 KDEL suggests that this form is more stable and that it might be retained intracellularly in the ER while the FL-AGR2-WT might be partially secreted even in its full-length form [180]. The very weak expression of mAGR2 KTEL suggests that this form is less stable or it is secreted. However, the secretion of AGR2 in the spent culture media was not immunoblotted to confirm these hypotheses. The two bands detected from N-term AGR2 KDEL could mean that AGR2 is cleaved possibly by a signal peptidase to remove its signal peptide aa 1-20 that is not made when it is matured and perhaps the signal peptide is fused to aa 41 which is not natural for the protein (due to the deletion aa Δ20-40). Note that N-term AGR2 KDEL in the weak promoter RSV, only one band was predominant (lower band) but in the stronger promoter CMV, the second upper bands were equally visible (Figure 4.13C, compare RSV and CMV). This may suggest that the strong promoter AGR2 can bypass the post-translational modification leading to the expression of full-length N-term AGR2 KDEL.

4.2.5.2. Cloning and expression of structural variants of untagged-EpCAM

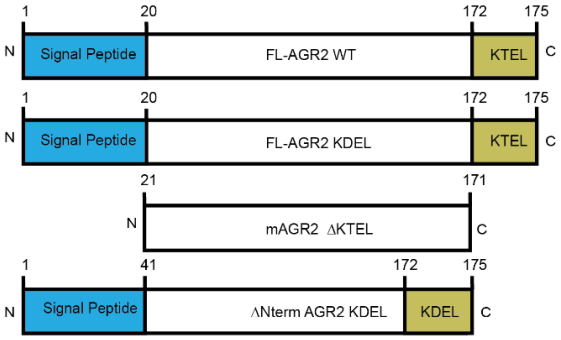
EpCAM also was cloned into the RSV and CMV promoter vectors so that our assay would be more accurate (i.e all genes were cloned into the same vectors). We cloned the full-length EpCAM (FL-EpCAM) and also the mature EpCAM (mEpCAM) which is without its signal peptide aa 1-23) (Figure 4.14A). These EpCAM cDNAs were also chemically synthesised harbouring *NcoI* and *XbaI* restriction sites at 5' and 3' ends respectively at each cDNA form and cloned into the multiple cloning sites containing the same restriction sites in pSF-RSV and pSF-CMV. Resulting recombinant EpCAM constructs and their description was designated in Figure 4.14B.

FLO-cells transfected with low FL-EpCAM expression vector showed very weak band while FL-EpCAM in strong promoter vector showed increased in

expression (Figure 4.14D) as detected with EpCAM antibody in western blot. However, the expression of mEpCAM in both low and high expression vectors was not detectable (Figure 4.14D). The different level expression of FL-EpCAM in the two expression promoters, again showed that the promoter can drive gene expression as expected. The FL-EpCAM showed multiple minor bands which correspond to cleavage of intracellular and extracellular cleavage consistent with previous studies [214, 229, 252]. mEpCAM was unable to express suggest that it is not stable and suggest that the signal peptide is crucial for EpCAM maturation.

Although FL-EpCAM was shown to express in FLO-1, the expression is relatively low as it needs high exposure for the band to be seen in western blot. As a backup, we construct a full-length EpCAM which was human codon optimised (Thermo Scientific) which was proven to increase EpCAM expression in bacteria. Codon optimised EpCAM was chemically synthesised and designed to be cloned into pCDNA3.1 (+) which is under the control of CMV promoter (Figure 4.14C). The expression of the construct (FL-EpCAM-CO) as transfected in FLO-1 cells showed an elevated EpCAM in western blot (Figure 4.14D). As such, we plan to omit the mEpCAM as it is not express and use either FL-EpCAM cloned into CMV vector or FL-EpCAM-CO for subsequent experiments.

A



B

AGR2 variants	Resulting plasmid		Description
	RSV	CMV	
FL-AGR2 WT	pAM1L	pAM1H	Full-length AGR2 wild type
FL-AGR2 KDEL	pAM2L	pAM2H	Full-length AGR2 KDEL (Thr 173 to Asp). Enhanced ER localisation.
mAGR2 ΔKTEL	pAM3L	pAM3H	Mature AGR2 (deletion aa 1-20 signal peptide) and ΔKTEL (deletion aa 172-175). Enhanced AGR2 secretion.
ΔNterm AGR2 KDEL	pAM4L	pAM4H	Deletion of N-term AGR2 intrinsically disorder region (Δaa 20-40) and ΔKTEL (deletion aa 172-175). Enhanced dimer.

C

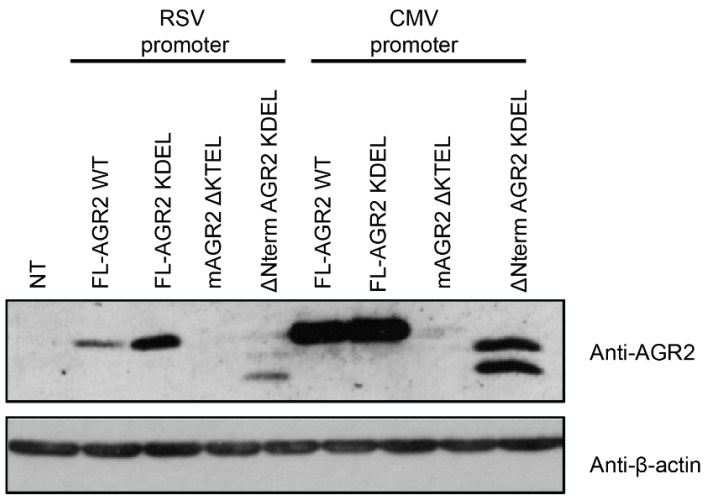
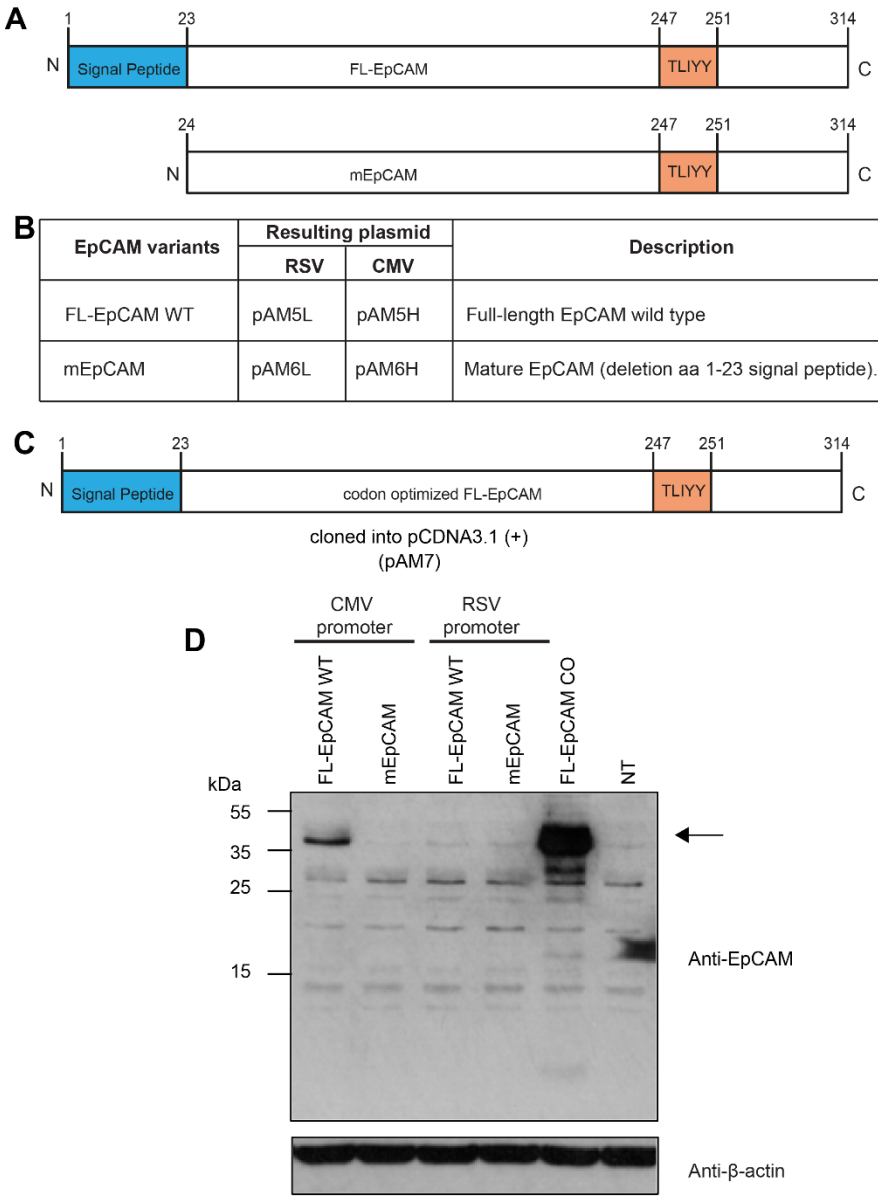


Figure 4.13 AGR2 constructs clone into weak and strong promoter vector and their respective expression in cells. (A) Schematic representation of different AGR2 variants clone into both weak (RSV) and strong (CMV) promoters vector for mammalian expression. AGR2 cDNAs were chemically synthesised containing NcoI at 5' end and XbaI at 3' end and cloned into a generic holding vector pMA-T (ThemoFisher Scientific). The cDNAs were then subcloned into RSV and CMV vectors using NcoI and XbaI. Plasmid names and description corresponding to different constructs were shown in (B). (C) FLO-1 cells were transfected with weak and high expressing AGR2 constructs and analysed by western blot using AGR2 polyclonal antibody K47. β-actin was used as loading control.



EpCAM variants	Resulting plasmid		Description
	RSV	CMV	
FL-EpCAM WT	pAM5L	pAM5H	Full-length EpCAM wild type
mEpCAM	pAM6L	pAM6H	Mature EpCAM (deletion aa 1-23 signal peptide).

Figure 4.14 EpCAM constructs clone into weak and strong promoter vector and their respective expression in cells (A) Schematic representation of different EpCAM variants clone into both weak (RSV) and strong (CMV) promoters vector for mammalian expression. EpCAM cDNAs were chemically synthesised containing *NcoI* at 5' end and *XbaI* at *3'* end and cloned into a generic holding vector pMA-T (ThemoFisher Scientific). The cDNAs were then subcloned into RSV and CMV vectors using *NcoI* and *XbaI.* Plasmid names and description corresponding to different constructs were shown in (B). (C) Schematic representation of wild-type full-length EpCAM which was chemically synthesised and human codon optimised for mammalian expression. This construct was designed to be cloned into a pcDNA3.1 (+) vector which under CMV promoter to drive gene expression. (D) FLO-1 cells were transfected with weak and high expressing or codon-optimized EpCAM constructs and analysed by western blot using EpCAM monoclonal antibody. β-actin was used as loading control.

4.2.6. Effect of AGR2 structural variants on EpCAM expression

Having successfully cloned and expressed the untagged-AGR2 and untagged-EpCAM, we reconstituted their expression in FLO-1 cells and see whether the structural variants of AGR2 have a negative or positive impact on EpCAM expression. We used weak (RSV promoter) and strong expression (CMV promoter) of AGR2 as means of low and high expressing AGR2, cotransfected with a constant amount of EpCAM without the need of titration. First, we tested the effect of AGR2 structural variants on FL-EpCAM/CMV and analysed in western blot. The low expression FL-AGR2-WT showed to elevate FL-EpCAM/CMV nearly two-fold as compared to transfected FL-EpCAM/CMV only (Figure 4.15A). However, the high expression FL-AGR2-WT did not significantly increase EpCAM expression. Similar results were observed for three other AGR2 forms, where the FL-AGR2-KDEL, mAGR2 ΔKTEL nor ΔNterm AGR2 KDEL significantly impact the expression of EpCAM/CMV (Figure 4.15A).

Next, we examined the effect of AGR2 structural variants on codon-optimized FL-EpCAM CO. Western blot analysis of co-transfected FLO-1 cells with AGR2 and codon optimised showed that all AGR2 variants in both low and high expression vector, elevate codon optimised EpCAM (Figure 4.15B). The enhanced ER form of AGR2 (FL-AGR2 KDEL) showed the most impact on EpCAM. The different level expression of AGR2 in all forms correlate to the expression of EpCAM; i.e the higher the expression of AGR2, more EpCAM expression was seen. The results obtained were different than the EpCAM cloned into CMV vector. We hypothesised that in the case of FL-EPCAM/CMV, only small amount of AGR2 is needed to achieve the FL-EpCAM/CMV steady-state levels while increasing the amount of AGR2 is toxic or there is a feedback mechanism to reduce the production of EpCAM. In the case of codon-optimized EpCAM, more AGR2 is needed to achieve the steady-state level of this form of EpCAM. Nonetheless, these data showed that AGR2, especially the FL-AGR2 WT when co-express with EpCAM in AGR2/EPCAM-null cells, was able to elevate the expression of EpCAM.

We also titrated the untagged FL-EpCAM CO in combination with AGR2-mCHERRY and co-transfected both constructs into FLO-1 cells. Western blot analysis showed that FL-AGR2-CHERRY but not mCHERRY was able to elevate the untagged EpCAM protein production Figure 4.16 which was similarly seen with EpCAM-GFP (Figure 4.10 and Figure 4.11). These data suggest that AGR2 functions in a positive role in EpCAM activity.

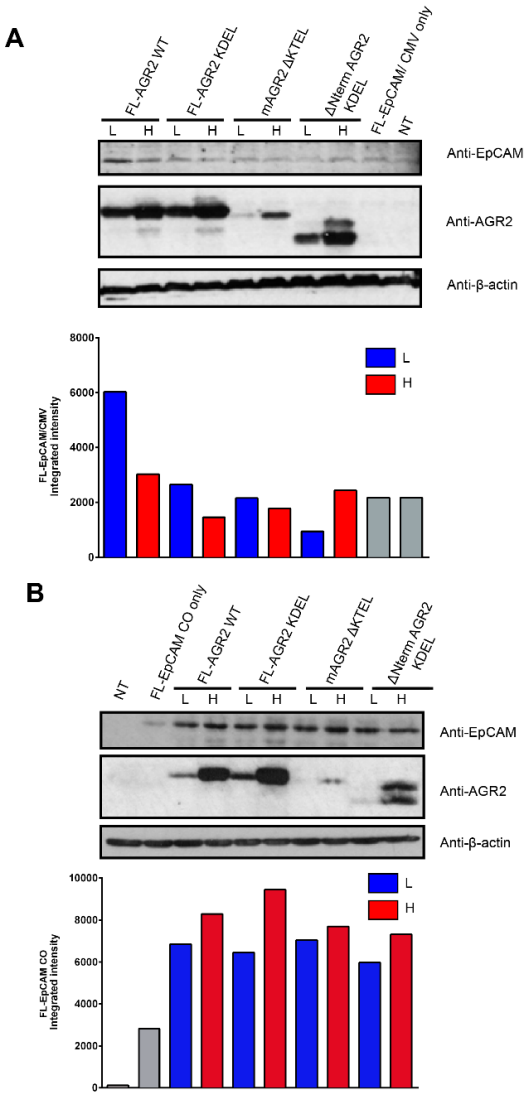


Figure 4.15 Effect of AGR2 structural variants on EpCAM expression. (A) (Upper panel) FLO-1 cells were transfected with FL-EpCAM/CMV alone or co-transfected with AGR2 structural variants (both low (L) and high (H) expression. An equal amount of lysates were electrophoresed and immunoblotted with anti-EpCAM to monitor the effect of EpCAM/CMV expression upon co-transfection with AGR2 constructs. The expression level of AGR2 constructs was blotted with anti-AGR2. β-actin was used as loading control. (lower panel) Integrated intensities correspond to FL-EpCAM/CMV bands were plotted in a bar graph **(**A) (Upper panel) FLO-1 cells were transfected with codon optimised EpCAM (FL-EpCAM CO) alone or co-transfected with AGR2 structural variants (both low (L) and high (H) expression. An equal amount of lysates were electrophoresed and immunoblotted with anti-EpCAM to monitor the effect of EpCAM/CMV expression upon co-transfection with AGR2 constructs. The expression level of AGR2 constructs was blotted with anti-AGR2. (Lower panel) Integrated intensities correspond to FL-EpCAM CO bands were plotted in a bar graph β-actin was used as loading control.

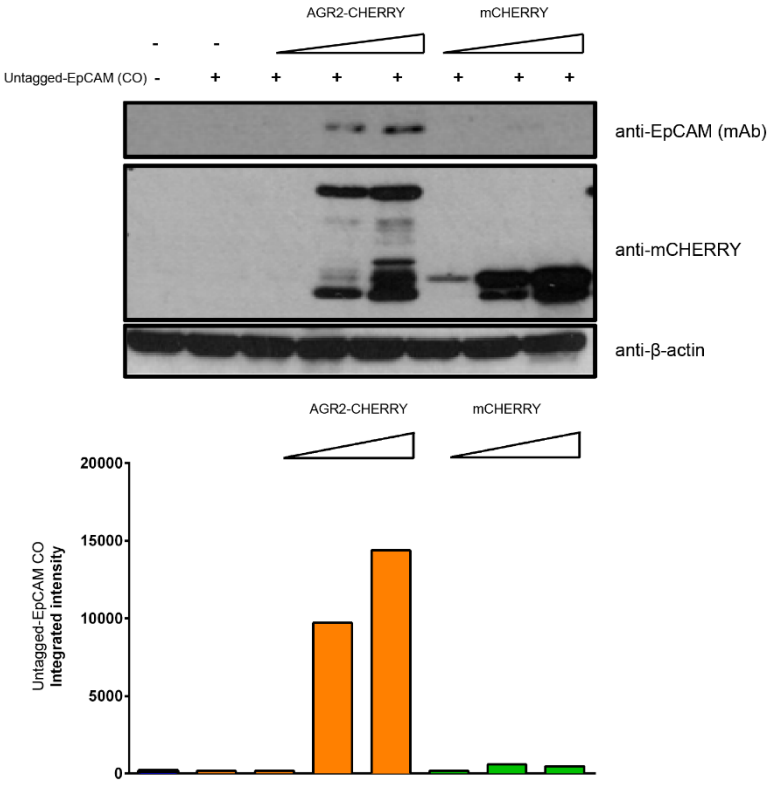


Figure 4.16 Western blot showing untagged FL-EpCAM CO expression in FLO-1 cells were elevated with titration of AGR2-CHERRY. (Upper panel) FLO-1 cells were either transfected with a low amount of untagged-EpCAM-CO (0.125 µg) alone or co-transfected with a fixed amount of EpCAM (0.125 µg) and in a combination with increasing amount of mCHERRY, AGR2-CHERRY (0.25 µg, 0.5 µg, 1 µg) and analysed using western blot with anti-EpCAM antibody to detect the presence of EpCAM. Only the band correspond to full-length AGR2-GFP was shown. The anti-mCHERRY antibody was used to detect the presence of mCHERRY, AGR2-CHERRY. β-actin was used as loading control. (Lower panel) Integrated intensity corresponding to untagged FL-EpCAM CO band was scored as a bar graph.

4.2.7. Immunohistochemistry using AGR2 and EpCAM antibody on oesophageal adenocarcinoma tissue microarray

Tissue microarray (TMA) is a high-throughput molecular pathology technique that allows assessment of biomarkers expression simultaneously on hundreds of multiple specimens at the same time [253]. It also allows simultaneous and parallel molecular profiling of clinical samples at the DNA, RNA, and protein level and also provides utmost preservation and use of limited collection tissue samples. TMA is produced by extracting microscopic tissue cores from different paraffin donor blocks and re-embedding these into a single recipient (microarray) paraffin block at defined array coordinates. First, the donor blocks containing invariably stored paraffin blocks are retrieved and sectioned to produce standard microscopic slides that are stained with hematoxylin and eosin. An experienced pathologist examines the slides to mark the area of interest (Cancer or normal tissues), after which the samples can be arrayed. Up to 1000 tissue samples can then be arrayed into a single paraffin block. This method gives several advantages for researchers [254]. First, it decreases assay volume as hundreds of biopsies can be assayed from a single microscope slide and in turn save time and cost. It also allows amplification of a limited tissue resource of biopsies and allows simultaneous analysis of a very large number of specimens as each tissue microarray paraffin block can be sectioned and cut multiple times and permits different individual assay from the same source. TMA also reduces experimental variation as each tissue sample in the array is treated in an equal manner and allow analysis of the entire cohort in one batch on a single slide. Thus, variables such as antigen retrieval, temperature, incubation times, washing procedure, and reagent concentration are consistent for the entire cohort. It is a practical and effective tool for high-throughput molecular analysis of tissues that is helping to identify new diagnostic and prognostic markers and targets in human cancers and has a range of potential applications in basic research, prognostic oncology and drug discovery.

We evaluated the expression of AGR2 and EpCAM and assess whether there is any evidence for the two proteins being co-expressed in human cancer clinical samples. A TMA containing 115 cores of oesophageal adenocarcinoma tissues was processed for immunohistochemistry using antibodies to AGR2 or EpCAM (Figure

4.17A). AGR2 and EpCAM were scored using a modified Allred method [255] which is based on the sum of intensity and area of staining (Figure 4.17B). Not all cores could be scored due to loss of material during experimental setup or lack of appropriate tissue type in the core. A total of 91 cores were intact and subjected to analysis. A summary of the expression of both AGR2 and EpCAM in the same tumour sample as a function of % cells per sample and intensity per sample was shown in Figure 4.17C. The scores showed that majority of the tissues stained for AGR2 and EpCAM. Nearly half of the populations, 46% (42/91) showed the highest expression (score 7) of EpCAM and AGR2. Tissues with at least a total of score 2 for EpCAM or AGR2 which we deemed to be the minimal score for positive staining were 90% (82/91) of the samples. This suggests that AGR2 and EpCAM were highly expressed in clinical tissue samples such as oesophageal adenocarcinoma. Only 9% (9/91) of samples exhibited no detectable expression of either protein and only one cancer exhibited high EpCAM expression with no detectable AGR2 protein expression. There is also a significant positive relationship between EpCAM expression and AGR2 expression (Pearson's correlation coefficient $r(89)=.8204$, $p<0.0001$) suggesting that the two proteins are highly co-expressed in oesophageal adenocarcinomas. Representative images of immunohistochemistry using AGR2 and EpCAM in TMA containing oesophageal adenocarcinoma were shown in Figure 4.18. AGR2 and EpCAM showed relatively similar staining coverage of staining as shown in representative images in Figure 4.18. Since the TMA cores were derived from the same samples but with different cuts, the same staining coverage of EpCAM and AGR2 suggests that the two proteins are co-expressed in most tissues of oesophageal adenocarcinoma.

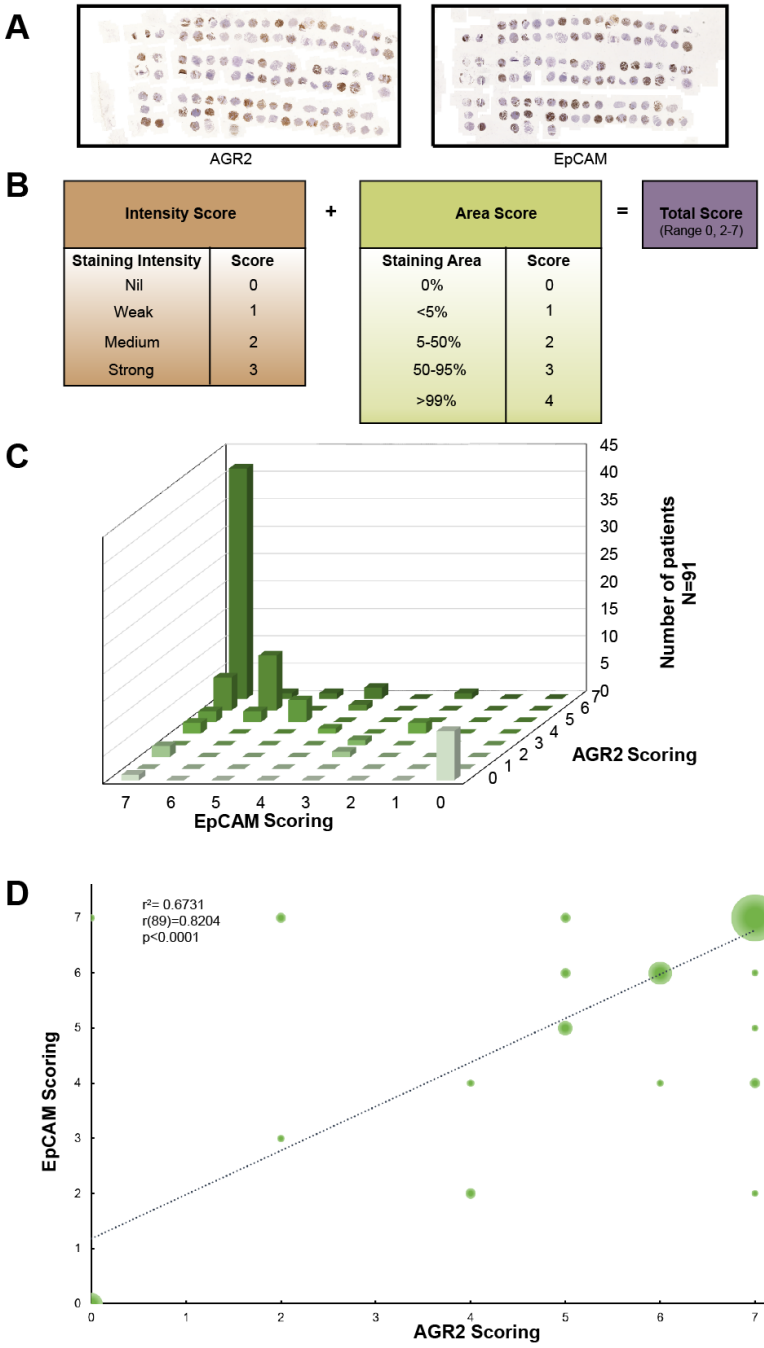


Figure 4.17 Immunohistochemistry of tumour microarray (TMA) containing oesophageal adenocarcinoma tissue cores using EpCAM and AGR2 antibody. (A) Image of a slide containing sections of TMA containing 115 cores of oesophageal adenocarcinoma tissue stained using antibodies to AGR2 (left) or EpCAM (right) (B) Scoring method for IHC using a modified Allred method [255]. Intensity was graded from 0-3 and area of staining was scored from 0-4. The sum of the intensity and area was calculated for each scorable core. (C) Bar graph summary of the expression of both AGR2 and EpCAM in the same tumour sample as a function of area and intensity per same sample. Only 91 out of 115 cores could be scored due to loss of material during IHC or lack of tissue type in the core. (Samples preparation and staining were performed by Dr Robert O'Neill).(D) Regression line showing a correlation of AGR2 and EpCAM scoring in the same tumour sample. Data were analysed using Pearson's correlation coefficient r(89)=.8204, $p<0.0001$. The size of the bubble indicates number of occurrences of the scoring.

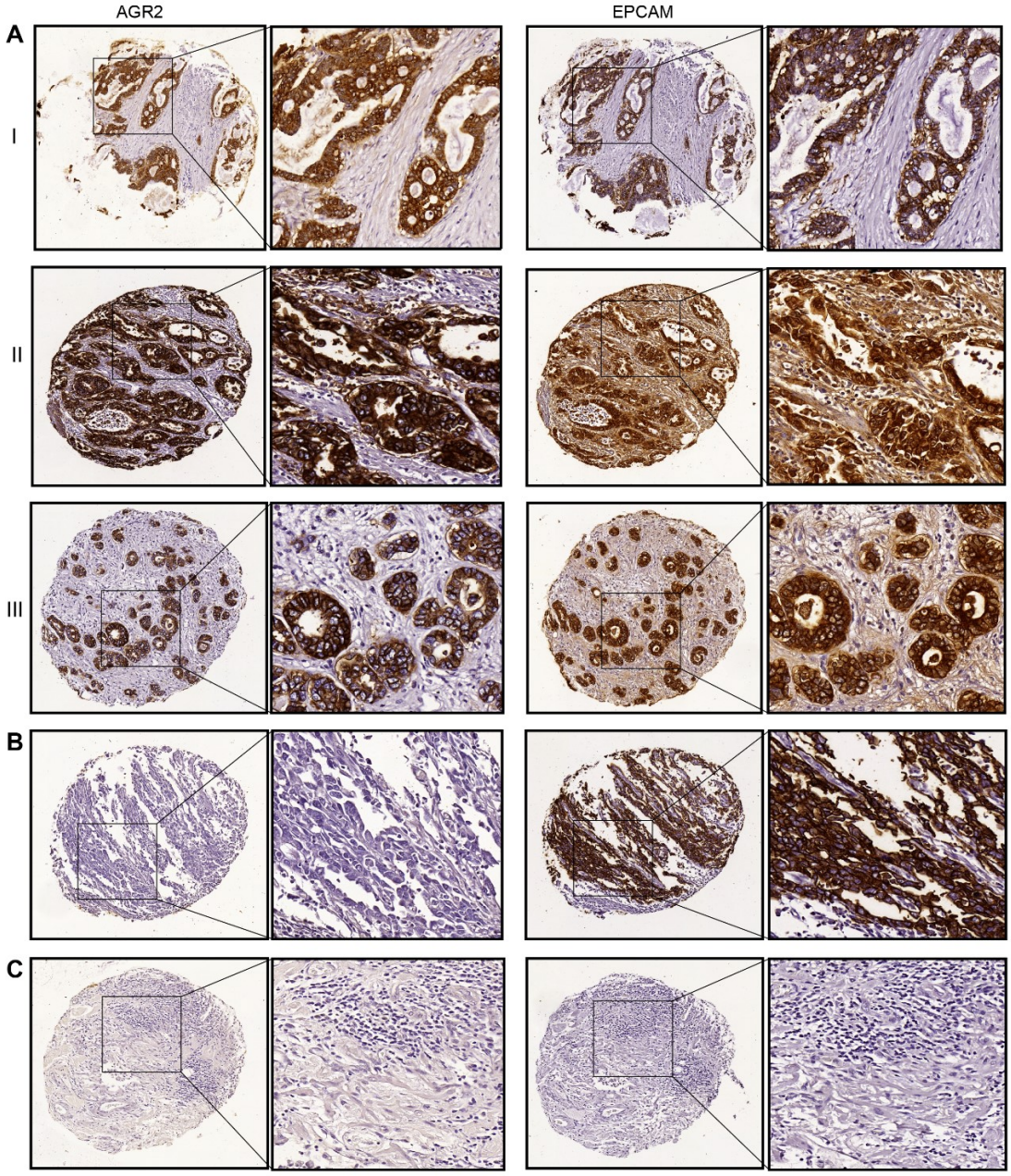


Figure 4.18 Representative images of immunohistochemistry using AGR2 and EpCAM in TMA containing oesophageal adenocarcinoma. (A) I-III Three representative images of typical high-level expression of both AGR2 and EpCAM. All three images have the highest score (7) of AGR2 and EpCAM. (B) Image of high-level of expression of EpCAM but no detectable AGR2 expression. (C) Image of no detectable expression of both proteins.

4.3. Discussion

Functions of AGR2 are not well defined since AGR2 is not only found in ER but also in other subcellular localisation such as nucleus [81, 83], cell surface and extracellular space [100]. AGR2 can form mixed disulphide bonds with transmembrane protein MUC2, in which AGR2 play a role in mucin processing [60]. AGR2 can interact with the metastasis-associated GPI-anchored C4.4a protein and extracellular alpha-dystroglycan (DAG-1) [58] and membrane receptor Prod1 [54]. Besides, AGR2 was found to be secreted during the development of pancreatic cancer [143] and was detected in the urine of prostate cancer patients [152] as well as in the blood of ovarian cancer patients [135]. A recent study describes the physiological role of extracellular AGR2 (eAGR) using organoid model derived from human lung epithelial cells. The study highlights that eAGR2 is a novel and unexpected microenvironment regulator in epithelial tumorigenesis and morphogenesis [180]. Altogether, this suggests that AGR2 can sculpt the cellular secretome. We have shown that AGR2 protein has a unique property for a protein disulphide isomerase in that it can bind sequence-specifically to a TxIL[YF][YF] peptide-docking motif present in client proteins such as receptor EpCAM in the previous chapter. Transmembrane/receptor proteins are the predominant class of proteins with such a motif suggesting that their maturation depends on AGR2 signalling. In this chapter, we further characterised the AGR2: EpCAM interaction using cell-based assays to extend the understanding how the motif on EpCAM (aa 247-25) facilitates its maturation or trafficking. Additionally, protein-protein interaction *in-vivo* may involve large, multidomain, and complex, and undergo many posttranslational modifications, which may influence maturation process like AGR2:EpCAM interaction [256].

First, by using PLA in MCF-7 cells, AGR2 and EpCAM can form complex in cells which are represented by PLA punctate foci (Figure 4.3A) but not in FLO-1 null cells (Figure 4.3B). Second, we used fluorescently tagged AGR2 and EpCAM for co-localisation assays in fixed cells. Paraformaldehyde fixation was used in this study as it has been widely used to preserved cell morphology prior to immunofluorescence staining. Fixed cells are no longer living which may create artefacts due to the disruption of cellular structures or lowering of protein antigenicity [257] and this may

lead to misinterpretation of localisation. However, using this methodology, GFP-tagged wt-EpCAM for example can localise to the plasma membrane, where it has been published to reside [258]. These data indicate that the GFP-tag does not preclude membrane localisation of EpCAM. The advantage of this GFP-tag expression assay was that we could also mutate EpCAM at Y251 to create an attenuated AGR2-interacting mutant that in turn demonstrated its' mislocalisation in cells. The disadvantage of the GFP-tag expression methodology is that it relies on using tagged proteins that might alter protein functions. Upon transfection of AGR2-CHERRY into cells, we discovered that there are some cleavages of the mCHERRY tag but not on GFP tag on EpCAM (Figure 4.4). The cleavage of mCherry has been documented in other studies as well. For example, one study showed that C-terminal mCherry fusion protein showed significant proteolytic cleavage [259] though the full-length mCHerry fusion protein can still be obtained. The studies also showed that different linkers connecting the parent protein and the fluorescent tag also may affect cleavage suggesting the instability of the mCherry tag. In addition, the addition of mCherry at the C-terminal of AGR2 (after KTEL sequence) might also affect protein targeting. The addition might obscure the recognition by KDEL receptor for retrieval from Golgi apparatus by retrograde transport to the ER lumen. A recent study showed that the addition of extra residues after KTEL sequence causes AGR2 to be secreted into cell media [68], suggesting that the addition of sequences upstream the ER retention sequence can disrupt protein localisation which may be due to KDEL receptor fail to recognise KTEL motif. We think that it would be better in the future if the construct is designed with the mCherry tag is placed before the KTEL for proper protein localisation. However, to do this, multiple sequential cloning needs to be performed. Since the same construct has been used previously used [260] (though with C-terminally DsRed tag) by our lab, and the fact that our construct produced active full-length AGR2 fusion protein as shown in western blot and cytoplasmic staining by fluorescence microscope, we think that this construct is still relevant to use in this studies. Co-transfection of AGR2-CHERRY and EpCAM-GFP proteins into MCF-7 cells showed that AGR2 localisation is cytosolic (Figure 4.5C) while EpCAM is largely plasma membrane bound (Figure 4.5D). Upon co-transfection of AGR2-CHERRY and EpCAM-GFP, a degree of colocalisation of the two proteins is observed

in the cytosol, suggesting that the two proteins can partially form stable interactions *in-vivo* (Figure 4.6). This finding is consistent with the PLA data where the foci seen in MCF-7 cells are also cytosolic which suggest that AGR2 interactions are confined to prior events in the maturation of the membrane-bound EpCAM. We next evaluated whether the TLIYY motif on EpCAM plays a role in its localisation and/or binding to AGR2 in MCF7 cells. If so, this would place more significance on its physiological relevance for further study.

Mutation of the peptide-docking motif loss-of-function, EpCAM^{Y251A}-GFP was not expressed at the plasma membrane when transfected into MCF7 cells (Figure 4.5E). Upon AGR2-CHERRY co-transfection, the EpCAM^{Y251A}-GFP mutant remained localised predominantly to the cytosol and/or nuclear membrane and did not exhibit a distribution of wt-EpCAM-GFP (Figure 4.6B). These data suggest that this protein-protein interaction pair has impacts on each other in cell systems. Thus, we conclude from the cell-based validation that, although that there is not a strikingly "stable" co-localisation of AGR2-CHERRY and EpCAM-GFP in every co-transfected cell, the key results are: (i) a degree of "transient" co-localisation of AGR2-CHERRY and EpCAM-GFP and (ii). Mislocalisation of EpCAM^{Y251A}-GFP suggests that the AGR2 docking site is important, in cells, for the appropriate trafficking of EpCAM to its destination.

This chapter showed that AGR2 act as a regulator of EpCAM signalling by promoting receptor presentation from the endoplasmic reticulum to the cell surface. It is possible that the docking site of AGR2 in EpCAM can stimulate EpCAM glycosylation for its plasma membrane representation. However, based on mobility in an SDS-PAGE, EpCAM stimulated for delivery to the plasma membrane in FLO-1 cells is in the non-glycosylated state, suggesting that AGR2 interaction with EpCAM occurs prior to EpCAM post-translational modification or without the help of glycans. This is consistent with our previous data in Chapter 3 which showed that AGR2 binds to a non-glycosylated form of EpCAM which takes place in the ER before it is released for subsequent maturation steps in the secretory pathway. This study also presents a mechanism for controlling EpCAM-mediated signalling. We demonstrate that AGR2

is able to dictate signalling activity by determining whether EpCAM is delivered to the plasma membrane from the secretory pathway. We reconstituted EpCAM expression in FLO-1 cells which are negative for both AGR2 and EpCAM expression. Transient transfection of FLO-1 cells with recombinant EpCAM-GFP alone resulted in an intracellular distribution (Figure 4.8). Upon co-transfection with EpCAM-GFP and AGR2-CHERRY cDNA, EpCAM-GFP cell surface expression is greatly stimulated (Figure 4.8 and

Figure 4.9). AGR2 has been shown to regulate another membrane-bound protein namely EGFR [111]. Similarly, EGFR-mediated signalling was shown to be dependent on AGR2 expression and is required for receptor delivery to the plasma membrane and thus EGFR signalling. However, the study showed that the AGR2 acts a novel post-translational regulator of EGFR-mediated signalling which suggests that AGR2 can escape ER and has different roles in regulating membrane bound proteins. We also found that AGR2 can elevate the expression of EpCAM protein by chance. Titration of AGR2-CHERRY with fix amount EpCAM-GFP co-transfected into FLO-1 cells increase the fluorescence staining of EpCAM-GFP (Figure 4.10) and the increased level of EpCAM protein is confirmed by western blot (Figure 4.11). Based on these findings, we hypothesise that AGR2 may serve as a regulatory factor, if not major, in EpCAM-mediated signalling. EpCAM molecule has been targeted for cancer immunotherapy such as Catumaxomab [261] which target cell surface EpCAM. Therefore, EpCAM cell surface representation is necessary for the drug recognition which can be achieved by the AGR2 expression, so any manipulation of AGR2 expression is relevant for the effectiveness of current therapy. Current data can be further strengthened to show that EpCAM signalling is dependent on AGR2 expression. For example, FACS experiment can be used to count the number of cells with EpCAM membrane staining in the presence or absence of AGR2. This can be done by staining cells with antibodies recognising extracellular EpCAM (which is used in this study) on unfixed FLO-1 cells in the presence and absence of exogenous AGR2. In addition, it would be interesting in the future to deplete AGR2 by short-hairpin RNA (shRNA) or knockout AGR2 cell models which can then be used to check whether EpCAM cellular distribution is affected by the altered level of AGR2 expression. It is

also important to establish the rate of EpCAM synthesis and maturation in the presence of absence of AGR2 which can be done by radiolabel pulse-chase experiment.

Protein folding and assembly in the ER is largely facilitated by chaperones and devoted ER-resident enzymes like the heat shock proteins and protein disulphide isomerases [262]. AGR2 belongs to protein disulphide isomerase superfamily containing thioredoxin domain. This domain is responsible for interaction of AGR2 client proteins through the formation of mixed disulphide bonds which were seen in MUC2 [60] and EGFR [111]. Mutating the single cysteine within the thioredoxin domain, C81S still showed stimulation of EpCAM-GFP cell surface expression, suggesting that a chaperonin function rather than a redox function is directly responsible for EpCAM elevation (Figure 4.11).

Proteins destined for conventional secretory pathway enters ER and undergo stringent quality control before they can proceed to the Golgi apparatus, and subsequently to the final destination [256]. EpCAM contains N-terminal signal peptide which suggests that the protein is targeted to the ER and secretory pathway. Following the ER quality control (ERQC), correctly folded EpCAM is then able to exit the ER and subsequently to the plasma membrane where signalling is transduced [256]. The ability of AGR2 to transport membrane-bound EpCAM suggests that it impacts cell signalling. Although generally reside in the ER, AGR2 also can leak through the cytosol or is secreted, suggesting that AGR2: EPCAM interaction can occur outside the ER. Mutating the C-terminal ER retention site on AGR2 to a more canonical ER retention motif (KTEL to KDEL), showed a slight increase in stimulating (compared to wt-AGR2 stimulation) the production of codon-optimized EpCAM (Figure 4.15B) implying that protein-protein interaction occurs inside the ER for proper folding and assembly of EpCAM. We also cloned various forms of AGR2 based on previously published data to check whether any of these AGR2 variants on EpCAM signalling (Figure 4.13). Based on the western blotting experiment, these AGR2 variants have minimal impact on EpCAM expression (Figure 4.15). Nevertheless, these constructs can be further manipulated for future cell-based assays.

A recent finding by our group using quantitative shotgun proteomics in oesophageal adenocarcinomas vs. normal tissues biopsies, have identified EpCAM as one of the most independent validated "target" in the tumour sample (Dr Robert O'Neill Thesis, and manuscript in preparation). Immunohistochemistry on a TMA containing 91 cores of oesophageal adenocarcinoma tissues in this study confirmed that EpCAM is highly expressed (Figure 4.17). Similarly, AGR2 was also found to be highly expressed in the vast majority of this type of adenocarcinomas. Interestingly, tissue samples with the maximum scoring of EpCAM and AGR2 are represented by nearly half of the population (Figure 4.17). Another set of immunohistochemistry analysis using normal squamous tissues showed that AGR2 expression was absent but exhibited low expression in normal gastric. EpCAM expression, however, was observed in both normal squamous and normal gastric epithelia but the expression was relatively low compared to the expression in adenocarcinoma tissues (Dr Robert O'Neill Thesis, and manuscript in preparation). The staining pattern and coverage of AGR2 and EpCAM on the tissues suggest that both proteins are co-expressed in tumour samples at least in the case of oesophageal adenocarcinoma (Figure 4.18). The absence of AGR2 protein in most normal tissues but high expression in tumour tissues implies that AGR2 expression represents a novel and probably a dominant post-translational mechanism for regulating EpCAM-mediated cell signalling. There is also a significant positive relationship between EpCAM expression and AGR2 expression (Pearson's correlation coefficient $r(89)=.8204$, $p<0.0001$) suggesting that the two proteins are highly co-expressed in oesophageal adenocarcinomas (Figure 4.17). Since we have yet to receive the patient's data, we were unable to correlate with patient's cancer progression specifically patient's survival in which we will incorporate in near future.

In conclusion, we presented data showing AGR2 and EPCAM can interact in cells consistent with our *in-vitro* biochemical results in the previous chapter. We also developed a cell-based protein-protein interaction assay for studying AGR2 and EpCAM interaction by reconstituting their expression in FLO-1 cells (AGR2-/EpCAM-). Using this reconstituted complex, EpCAM delivery to the plasma membrane was shown to be dependent on AGR2 expression showing that AGR2 act

as a major regulator in EpCAM signaling Finally, AGR2 and EpCAM were highly expressed in vast majority clinical tissues and we provided data using oesophageal adenocarcinomas tissues that showing AGR2 and EpCAM was indeed highly expressed in this type of cancer and showed positive relationship. This suggests that both proteins may serve as physiologically relevant candidates for oesophageal adenocarcinoma biomarker for early diagnosis and for future therapeutic intervention.

CHAPTER 5: Analysis of AGR2 Reprogramming Cancer Cell Using Quantitative Proteomics

5.1. Introduction

5.1.1. Genetic screen to study gene function

The exploration of a gene often begins with a DNA sequence. As previously discussed, the main challenge is to translate the DNA sequence into a function or distinct phenotype. The standard genetic approach to decipher gene function is to disrupt normal gene expression and study the resulting phenotypes. This technique is called classical forward genetics which has been performed for more than 100 years [263]. This method determines gene that is responsible for a phenotype of an organism (i.e mutant phenotype → sequence). This was initially done by using naturally occurring mutations or inducing mutants with radiation, chemicals, or insertional mutagenesis using a virus. Subsequently, mutant individuals are isolated, and then the gene is mapped. However, this approach is painstaking due to numerous technical challenges to map random and superficially small changes in an enormous genomic DNA.

With large volumes of genomic sequence data that move rapidly in the past decades, many genetic sequences are discovered in advance without knowing their biological function. Therefore, instead of going from phenotype to sequence as in forward genetics, geneticist used so-called reverse genetics that works in the opposite direction, which determines the function of a gene by analysing the phenotypic effects of altered DNA sequences (sequence → mutant phenotype). This approach begins with a known gene and experimentally investigates the effects of altering the sequence or expression of that gene to infer gene function [264]. The alteration of the sequence can either be targeted specifically (e.g gene silencing or homologous recombination) or non-specifically (e.g., chemical mutagenesis, transposon-mediated mutagenesis) followed by screening a library of phenotypic individuals at a specific location.

5.1.2. New genome editing technologies to study gene function

Gene targeting by homologous recombination was first described by Capecchi et al. which integrates exogenous repair templates that contain sequence homology to the target site [266]. HR-mediated targeting has simplified the generation of isogenic animal models via manipulation of germline-competent stem cells that precisely alter target gene, advancing many types of biological research. However, the desired recombination events can be very challenging due to the low efficiency of a recombination event, presenting enormous challenges for large-scale applications of gene-targeting experiments [266] and generating knockin or knockout animal model can be very time-consuming. This, in turn, encouraged scientists to develop new tools for reverse genetics.. Reverse genetics complete loss-of-function approaches became available with the discovery of programmable nucleases which enable precise genome editing by introducing DNA double-strand breaks (DSBs) at specific genomic loci. DSBs subsequently recruit endogenous repair machinery for either non-homologous end-joining (NHEJ) or homology-directed repair (HDR) to the DSB site to mediate genome editing[265]. To date, four major classes of nucleases have been developed for sequence-specific genome editing: i) meganucleases and their derivatives, ii)zinc finger nucleases (ZFNs) iii) transcription activator–like effector nucleases (TALENs) and iv) (clustered regularly interspaced short palindromic repeats-CRISPR associated) CRISPR/Cas9 [266].Since these nucleases differ in their mode of DNA recognition, they can be generally classified into two categories: i) ZFNs, TALENs and meganucleases achieve specific DNA binding via protein-DNA interactions, whereas ii) Cas9 is targeted to specific DNA sequences by a short guide RNA molecule that base-pairs directly with the target DNA and by protein-DNA interactions. Description of the genome editing tools is summarised in Table 5.1

Once the DSB has been made, the lesion may be repaired by either one of the two major DNA repair machinery: Non-homologous end joining (NHEJ) or homology-directed repair (HDR). NHEJ may repair the lesion by direct re-ligation of the two DSB ends in a process that does not require a repair template. Although NHEJ-mediated DSB repair can be accurate, repeated repair of the same DSB by NHEJ

machinery eventually results in the formation of small insertion/deletion (indels) mutations bridging the break site. Indels introduced into the coding sequence of a gene can cause frameshift mutations or premature stop codon that lead to mRNA degradation by nonsense-mediated decay or result in the production of nonfunctional truncated protein, hence a gene knockout [267]. Alternatively, DSB can be repaired by the high-fidelity HDR, by introducing exogenous DNA repair template that can mediate the repair and precisely edit the target gene. Thus, NHEJ and HDR may be used to suppress gene function similar to RNAi, but it may lead to permanent suppression by introducing loss-of-function mutations into the genome of targeted cells.

Together, these powerful tools offer a new assurance to decipher any gene's function in a rapid and efficient way. These programmable nucleases have been shown to achieve effective genome editing in a wide range of model organisms and mammalian cells, and investigators are now focusing on to developing these tools as therapeutics [268, 269].

	ZINC FINGER NUCLEASE	TALEN	CAS9	MEGANUCLEASE
RECOGNITION SITE	Typically 9–18 bp per ZFN monomer, 18–36 bp per ZFN pair	Typically 14–20 bp per TALEN monomer, 28–40 bp per TALEN pair	22 bp (20-bp guide sequence + 2-bp protospacer adjacent motif (PAM) for *Streptococcus pyogenes* Cas9); up to 44 bp for double nicking	Between 14 and 40 bp
SPECIFICITY	Small number of positional mismatches tolerated	Small number of positional mismatches tolerated	Positional and multiple consecutive mismatches tolerated	Small number of positional mismatches tolerated
TARGETING CONSTRAINTS	Difficult to target non-G-rich sequences	5 targeted base must be a T for each TALEN monomer	Targeted sequence must precede a PAM	Targeting novel sequences often results in low efficiency
EASE OF ENGINEERING	Difficult; may require substantial protein engineering	Moderate; requires complex molecular cloning methods	Easily re-targeted using standard cloning procedures and oligo synthesis	Difficult; may require substantial protein engineering
IMMUNOGENICITY	Likely low, as zinc fingers are based on human protein scaffold; FokI is derived from bacteria and may be immunogenic	Unknown; protein derived from *Xanthamonas* sp.	Unknown; protein derived from various bacterial species	Unknown; meganucleases may be derived from many organisms, including eukaryotes
EASE OF *EX VIVO* DELIVERY	Relatively easy through methods such as electroporation and viral transduction	Relatively easy through methods such as electroporation and viral transduction	Relatively easy through methods such as electroporation and viral transduction	Relatively easy through methods such as electroporation and viral transduction
EASE OF *IN VIVO* DELIVERY	Relatively easy as small size of ZFN expression cassettes allows use in a variety of viral vectors	Difficult due to the large size of each TALEN and repetitive nature of DNA encoding TALENs, leading to unwanted recombination events when packaged into lentiviral vectors	Moderate: the commonly used Cas9 from *S. pyogenes* is large and may impose packaging problems for viral vectors such as AAV, but smaller orthologs exist	Relatively easy as small size of meganucleases allows use in a variety of viral vectors
EASE OF MULTIPLEXING	Low	Low	High	Low

Table 5.1 Comparison of different programmable nuclease platforms. (Reproduced from [266])

5.1.3. Aims of chapter

The aim of this chapter is to employ protein knockdown strategy to study AGR2 function. Our primary focus is to employ current genome editing technology to knockout AGR2 in a cell line followed by quantitative proteomics screen to identify dominant proteins, subcellular components and pathways affected by AGR2 depletion. This chapter serves as a second paradigm to identify potential AGR2 client proteins and to define the AGR2 regulatory network. However, our attempts to knockout AGR2 using the CRISPR/Cas9 gene editing tool was not successful. This led to a different strategy in which we utilised protein overexpression strategy to study AGR2 function. We utilised a AGR2-negative cell line that we identified earlier (Chapter 3) in which we reconstitute the cell as a function of AGR2 transient overexpression followed by quantitative proteomics screen. We have shown in Chapter 3 that AGR2-specific peptide motif dominantly exists in membrane proteins suggesting that AGR2 has a tendency to bind to this class of proteins. We have established that AGR2 can bind and mature one protein of this class represented by EpCAM using both in-vitro (Chapter 3) and cell based assays (Chapter 4), and showed that AGR2 act as a major regulator in EpCAM signalling by promoting receptor presentation from the endoplasmic reticulum to the cell surface. Thus, this chapter was also aimed to check whether proteomic screens showed large perturbation of membrane proteins or those pertaining to secretion and trafficking to support our working hypothesis that AGR2 could largely bind to and help in the maturation of membrane proteins destined for secretory pathway.

5.2. Results

5.2.1. Establishing AGR2 knockout isogenic cell line using genome editing CRISPR/Cas9

5.2.1.1. CRISPR/Cas9 technology

CRISPR-Cas is an adaptive immune system that uses RNA-guided nucleases to cleave exogenous genetic elements and is found in roughly 50% of bacteria and 90% of archaea [270]. There are three types (I-III) of CRISPR systems found in bacterial and archaeal hosts[268]. The Type II CRISPR system derived from *Streptococcus pyogenes* is one of the best and widely used. This system consists of the nuclease Cas9, the CRISPR RNA (crRNA) array that encodes the guide RNAs and a required auxiliary trans-activating crRNA (tracrRNA) that facilitates the processing of the crRNA array into discrete units. Each crRNA unit then contains a 20-nt guide sequence and a partial direct repeat, where the former directs Cas9 to a 20-bp DNA target via Watson-Crick base pairing (Figure 5.1A). The crRNA and tracrRNA can be fused together to create a chimeric, single guide RNA (sgRNA). The target DNA must possess a 5'-NGG protospacer adjacent motif (PAM). The PAM is a critical feature of the Cas9 system which dictates Cas9 nuclease to DNA target and mediates DSB at the 3' upstream of the PAM (Figure 5.1A). Cas9 nucleases can be mutated and exploited for an additional function such as to carry out strand-specific cleavage. For example, an aspartate-to-alanine (D10A) mutation in the RuvC catalytic domain allows the Cas9 nickase mutant (Cas9n) to nick rather than cleave DNA to produce single-stranded breaks. With the introduction of HDR, a more precise gene editing can be performed. A pair of sgRNA with Cas9n targeting each double-stranded DNA can be used simultaneously to mediate DSB, thus increasing the specificity of target recognition. Gene editing occurs during the DSB repair. The DSB at a target genomic locus typically undergoes one of two endogenous DNA repair mechanisms: Non-homologous end joining (NHEJ) or the homology-directed repeats (HDR) as described before (Figure 5.1B).

5.2.1.2. Guide RNA design for AGR2 knockout

The specificity of the Cas9 nuclease is determined by the 20-nt guide sequence within the sgRNA.. Since CRISPR/Cas9 is relatively new, we decided to outsource the

designing and cloning of gRNA for knocking out AGR2 to a biotechnology company Horizon Discovery[14].

We decided to use sgRNA-expressing plasmids to minimise experimental error as the gRNA is inserted within the same plasmid with the remainder of the sgRNA scaffold immediately following Cas9, and this only requires a single transfection.The gRNA designed for AGR2 knockout was also designed for AGR2 knockin mutation (AGR2 C81S), where this mutation will have defective oxidoreductase activity on its sole cysteine within the thioredoxin domain. This AGR2 mutation is chosen because most PDI-like AGR2 interactions are thought to occur via the disulphide bridge through the cysteine. Thus, it is interesting to introduce this mutation into cells to comprehend what proteins are affected as a result of the C81S knocking following quantitative proteomics. As the knockin require gRNA to target DNA sequence and introduce DSB, therefore the knockin gRNA can also be used for the knockout experiment. However, for a knockin to take place, it requires an extra donor template containing the knockin mutation. The AGR2 knockout/knockin gRNA (gRNA218) targets exon 4 of AGR2 (as the DNA sequence containing the Cys81 is in exon 4). This gRNA scored the highest quality score (for targeting exon 4) based on Horizon Discovery design tool and contains the least off-targets compared to other gRNAs. Therefore, we chose this gRNA initially in an attempt to knockout AGR2. Horizon discovery cloned the gRNA into a GE6-002 plasmid. This plasmid contains a U6 promoter, expressing wild-type Cas9 which is under the control of CMV promoter. This plasmid also contains a 2A puromycin resistance gene to allow screening and selection of transfected cells.

[14] https://www.horizondiscovery.com/

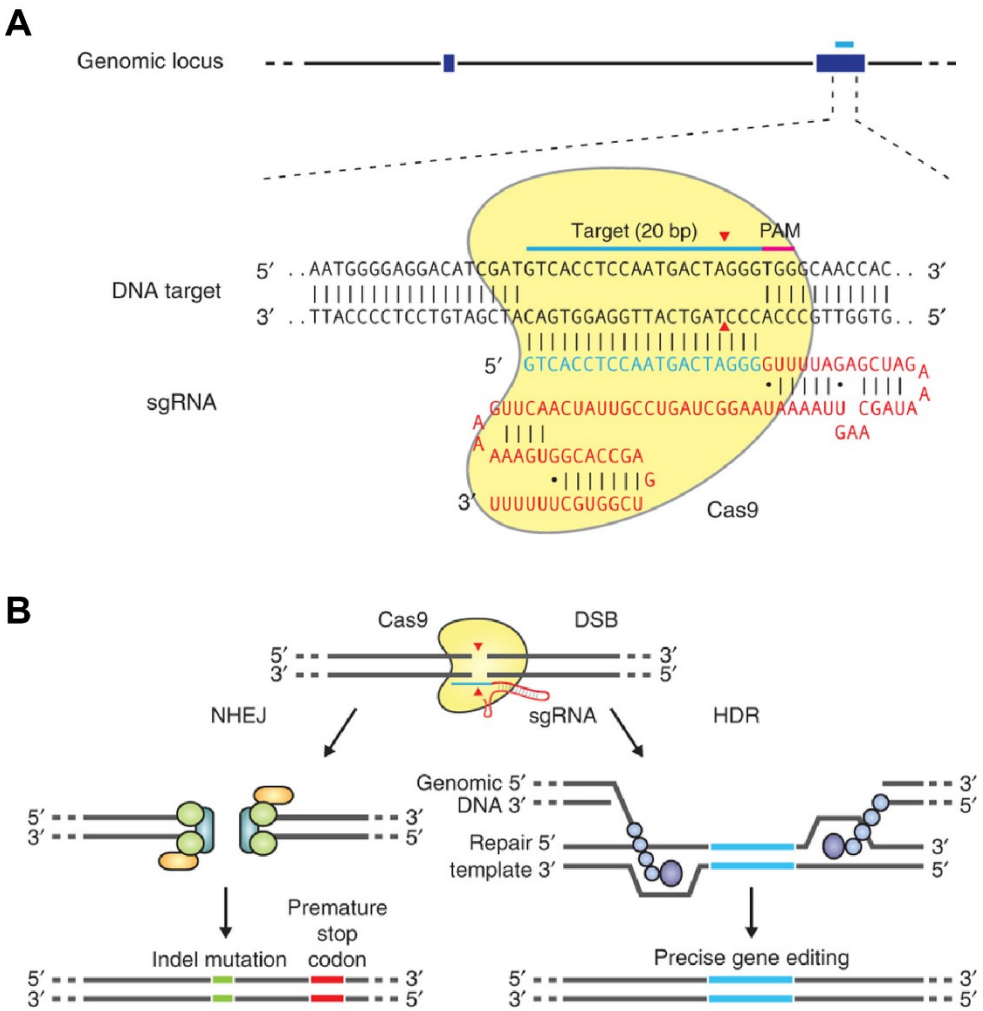


Figure 5.1 RNA-guided Cas9 nuclease and DSB repair that promotes gene editing. (From [271] (A) Schematic of the RNA-guided Cas9 nuclease. The Cas9 nuclease from *S. pyogenes* (in yellow) is targeted to genomic DNA) by a sgRNA consisting of a 20-nt guide sequence (blue) and a scaffold (red). The sgRNA sequence pair with the DNA target (blue bar on top strand), directly upstream of a requisite 5′-NGG adjacent motif (PAM; pink). Cas9 mediates a DSB ~3 bp upstream of the PAM (red triangle). B) DSBs induced by Cas9 (yellow) can be repaired in one of two DNA repair machinery. In the NHEJ pathway, the ends of a DSB are re-ligated, which can result in random indel mutations at the site of the junction. Indel mutations occurring within the coding region of a gene can result in frameshifts and the creation of a premature stop codon, resulting in gene knockout. Alternatively, HDR is another DNA repair machinery which can generate precise gene editing in the presence of an exogenously introduced repair template. Single-stranded nicks to the DNA can also induce HDR.

A

Product code	AGR2 exon	gRNA sequence	PAM	Orientation	No. off target	No. off target within genes
gRNA 218	4	CTTGATGATTATTCATCACT	TGG	Forward	27	3

B

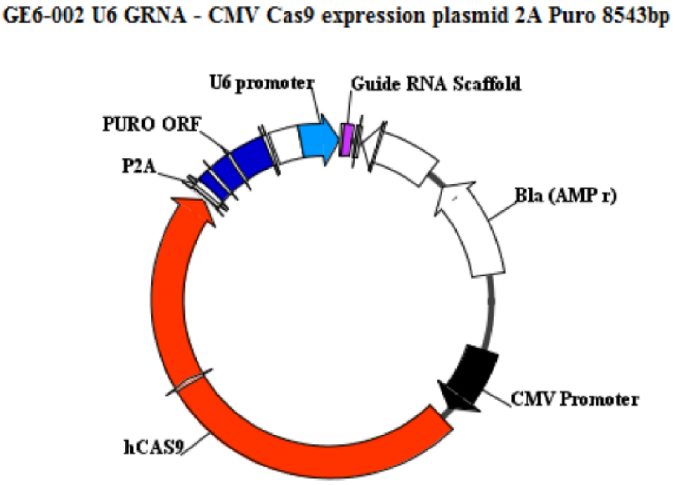


Figure 5.2 gRNA design for AGR2 knockout. (A) The 20-nt gRNA for AGR2 knockout is designed by Horizon Discovery. The gRNA targets the locus exon 4 of AGR2 genomic sequence. It as the lowest number of off-target binding. It can also be used to knockin C81S provided that the repair template is introduced alongside this gRNA. (B) The gRNA is cloned into GE6-002 which contains a U6 promoter, expressing wild-type Cas9 which is under the control of CMV promoter. This plasmid also contains a 2A puromycin resistance gene to allow screening and selection of transfected cells.

5.2.1.3. Experimental design for AGR knockout

We followed the general protocol suggested by Horizon discovery and the experimental design is summarised in Figure 5.3. We chose the breast cancer model cell line, MCF-7 as the target cell line where AGR2 protein expression level can be manipulated as used in previous experiments in Chapter 3-5. To deliver the plasmid into the cell line, we used transfection method using Attractene (Qiagen) as this is the routine transfection reagent used in our lab and was also proven to be the best compared to some commercial transfection reagents in term of cell viability and transfection efficiency. The transfected cells were incubated for 72 hours prior to single cell dilution into 96-well plates in standard cell culture media to obtain isogenic clonal population. After several days, each well of 96-well plates were checked and marked if they contained a single colony, while wells containing more than one colony (which may contain contaminating clonal line) were excluded. The clonal lines were then incubated in the incubator for several weeks to expand as described in Materials and Methods followed by characterisation and validation of the isolated clones as discussed below.

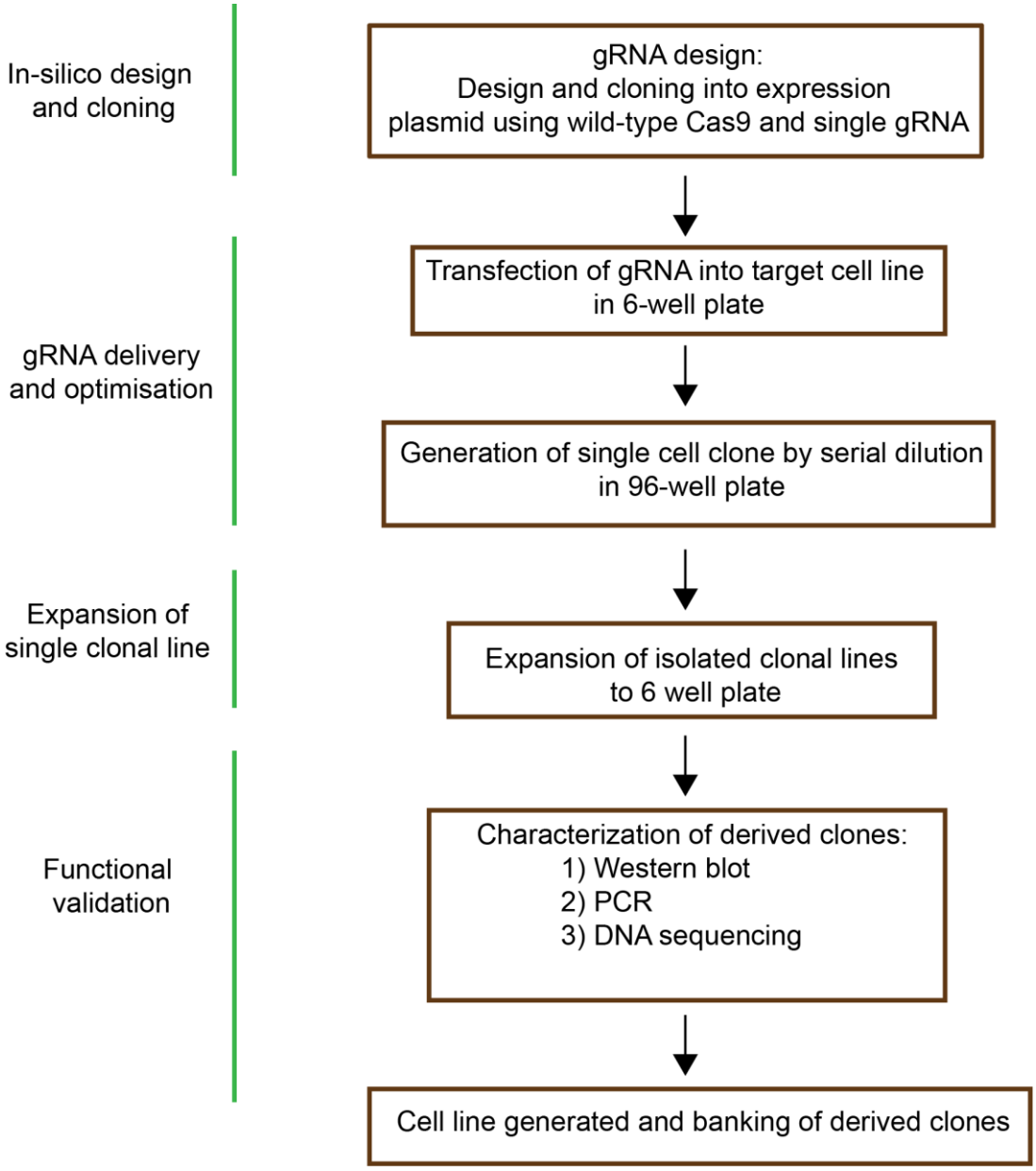


Figure 5.3 Experimental design for generating AGR2 knockout cell line.

5.2.1.4. Characterization of AGR2 knockout clones

To check whether the isolated clones carry AGR2 knockout, we initially used western blot to check AGR2 protein level as this is a quick and simple method to see the loss-of-function of a gene. If homozygous knockout is successfully carried out, there will be no detectable AGR2 band in western blot. Heterozygous knockout can also be obtained, where one of the AGR2 alleles are edited while the other allele is intact which may or may not affect the AGR2 protein expression level.

Upon expanding the clones, we noticed that the clonal lines have different growth levels, and these cells were harvested and labelled as followed: i) fast growth; ii) moderate growth iii) slow growth. Initially, more than 100 clones were successfully isolated to contain single colony, however, some cells died suggesting some clonal lines cannot tolerate single cell dilution. At the end, we only managed to isolate ~40 clonal lines. Western blot analysis of the AGR2 clonal lines using AGR2 antibody showed a variable degree of AGR2 expression (Figure 14b). One clonal line did not show any detectable AGR2 (Clone 11) that could potentially have an AGR2 knockout. Some clonal lines showed reductions of AGR2 (e.g Clones 15, 16, 19 and 20) compared to unedited non-transfected cells. This reduction of AGR2 expression could mean that the clones carry heterozygous knockout. Clonal lines with no detectable or reduced AGR2 expression were further grown and harvested for a second western blot analysis to make sure that their expression can be maintained. Second western blot analysis, however, showed that all clones including clone 11, our main candidate for AGR2 knockout showed AGR2 expression.

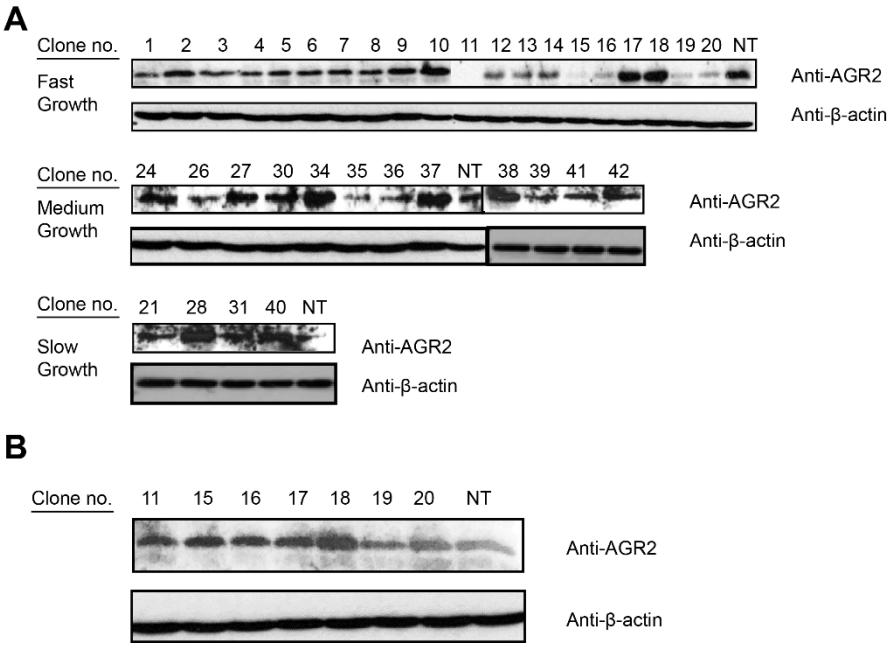


Figure 5.4 Western blot analysis of AGR2 knockout clonal line candidates. (A) Expanded clonal lines were harvested according to their cell growth levels: i) fast growth; ii) moderate growth iii) slow growth. The cells were harvested using urea lysis buffer and immunoblot using AGR2 K47 antibody. (B) The candidate clonal lines for AGR2 knockout were allowed to grow further and passaged several times before harvesting for a second western blot analysis using AGR2 K47 antibody. B-actin antibody was used as loading control.

5.2.1.5. Genotyping and sequencing of AGR2 knockout clonal lines

SURVEYOR nuclease assay is a method to detect a specific DNA mismatch caused by mutation, micro-deletion, inversion, or indels in heteroduplex DNA. This makes possible by the use of SURVEYOR nuclease, an endonuclease that cleaves DNA with high specificity at sites of base-substitution mismatch and other alterations. This method starts with the isolation of genomic DNA from the heterogeneous population of Cas9-targeted cells that is amplified by PCR. Amplicons are then denatured and re-hybridised slowly to generate heteroduplexes. The reannealed heteroduplexes containing indels are cleaved by SURVEYOR nuclease, whereas homoduplexes are left intact and these fragments can be resolved by DNA gel electrophoresis. Cas9-mediated cleavage efficiency (% indel) is then calculated on the basis of the fraction of cleaved DNA following analysis of integrated intensity of electrophoresed bands.

SURVEYOR primers for AGR2 knockout were designed from the genomic DNA region within AGR2 exon 4 that will give 586 bp long amplicon, where one of the primers is roughly one-third of the Cas9 target so that the cleavage fragments can be clearly distinguished in electrophoresis. SURVEYOR nuclease will recognise the indels occurring near Cas-9 cleavage site (3' bp PAM site) which will produce two more fragments; ~311 bp and ~275 bp in length. SURVEYOR nuclease assay requires a single PCR product to avoid non-specific cleavage during the digestion process. We examined the AGR2 SURVEYOR primers by performing PCR from the isolated genomic DNA of the non-transfected MCF-7 cells. DNA gel electrophoresis showed that the primers were able to produce a single PCR product with expected band size (586 bp). We next, extracted the genomic DNA from the AGR2 knockout clonal line candidates and assayed using SURVEYOR nuclease assay. Unfortunately, no cleavages were observed on all clonal lines suggesting that AGR2 knockout was not successfully carried out. Positive control was included in the assay kit where the DNA contains a single nucleotide polymorphism (SNP, G→C) in the reannealed heteroduplex DNA, which is recognized by the SURVEYOR nuclease, which produced the expected three bands (Uncut- 633bp*, G-G mismatch- 416 bp**, and C-C mismatch- 217 bp***). This also suggests that SURVEYOR nuclease works well

and is sensitive in detecting an SNP but not in the case of AGR2 knockout clonal line candidates. We further confirmed the results by analysing the PCR amplicons with DNA sequencing. The results showed that the genomic DNA of the clonal lines were intact with no alteration around the Cas9 target (3 bp upstream PAM site marked by red arrow) which confirm that the clonal lines did not carry AGR2 knockout.

We postulated that the gRNA has poor gRNA cleavage efficiency which can vary from cell line to cell line. To measure this, MCF-7 cells were transfected with gRNA218 and incubated for 72 hours before the cells were harvested for SURVEYOR assay. These transfected cells contain a pool of both edited and unedited genome which can be assayed by SURVEYOR nuclease. SURVEYOR nuclease assay on these mixed populations showed no cleavage of the genomic DNA on AGR2 exon 4 signifying that the sequence-specific guide gRNA218 did not direct the Cas9 nuclease to cleave the target locus. Another gRNA (gRNA219) provided by Horizon Discovery which showed the second best quality score to target AGR2 exon 4 (data not shown) was also assessed for its cleavage efficiency and the result was the same where no cleavage was seen. The combination of both gRNA218 and gRNA219 co-transfected into MCF-7 cells yield the same result. These data suggest that neither gRNAs were efficient for editing AGR2 gene. Though variable protein expression of AGR2 was seen in the clonal lines which may due to selection pressure from a single cell dilution and expansion.

A second attempt was made to knockout AGR2 by another student in the lab. This time, the gRNA was designed from the available CRISPR Design Tool[15] from the Feng Zhang lab at Massachusetts Institut of Technology (MIT) which was deemed to be the 'gold standard' at the time for designing gRNA. The MIT output colour codes the guides (green, yellow, red) designating high quality, mid quality and low quality. Guides which are categorised 'green' in the MIT output and that have the fewest off target sites specified in exons should be preferably chosen. The newly designed gRNA for AGR2 knockout targets exon 7 of AGR2 and has a high-quality score with minimum off-targets (Figure 5.7A). The gRNA was cloned in-house into

[15] http://crispr.mit.edu/

pLentiCRISPRv2 using BsmBI and the resulting clone was subjected to DNA sequencing to check the presence of the 20-nt gRNA (Figure 5.7B). We made some modifications in the experimental design (Figure 5.7C) where we used: i) Barrett's oesophagus cell line, CP-A which is positive for AGR2 and ii) fluorescence-activated cell sorting (FACS) to sort for single cell dilution. CP-A cells were chosen because they are routinely used in the lab and they are near-diploid cell line which could be manipulated relatively easier for gene editing. We also notice that our MCF-7 cell line is an aneuploid which maybe of hypertriploid or hypotetraploid which means that three of four copies of AGR2 needs to be depleted for complete homozygous knockout, which may also explain the difficulty of knocking out AGR2 in this cell line. For FACS single cell sorting, CP-A cells were co-transfected with gRNA construct and GFP-expressing plasmid (pmaxGFP, Lonza) and were sorted for GFP-positive cells. Following the expansion of the clonal lines and analysis by western blot, the results were disappointing as no cells produced any depleted AGR2 (data not shown). Again, like our first experimental layout, there were some clonal lines that showed reduced AGR2 expression, but SURVEYOR nuclease assay and DNA sequencing yield negative results as there were no evidence of genomic editing. This implied that there were selection pressures during single cell dilution and expansion that can give variable gene/protein expression levels. The inability to knockout AGR2 in two different cell lines with two different gRNA design sources might also suggest that knocking out AGR2 is lethal for the cells. Another explanation would be that the gRNA might have successfully directed Cas9 nuclease to introduce DSB in target genes but due to lethality, these edits cannot be assayed. Another possibility is that the gRNA bind to off-target sites (e.g gRNA218 has 30 off targets). However, these off-target sites were not checked because of time constraint (i.e primers need to be designed for each off-target for sequencing or for SURVEYOR assay).

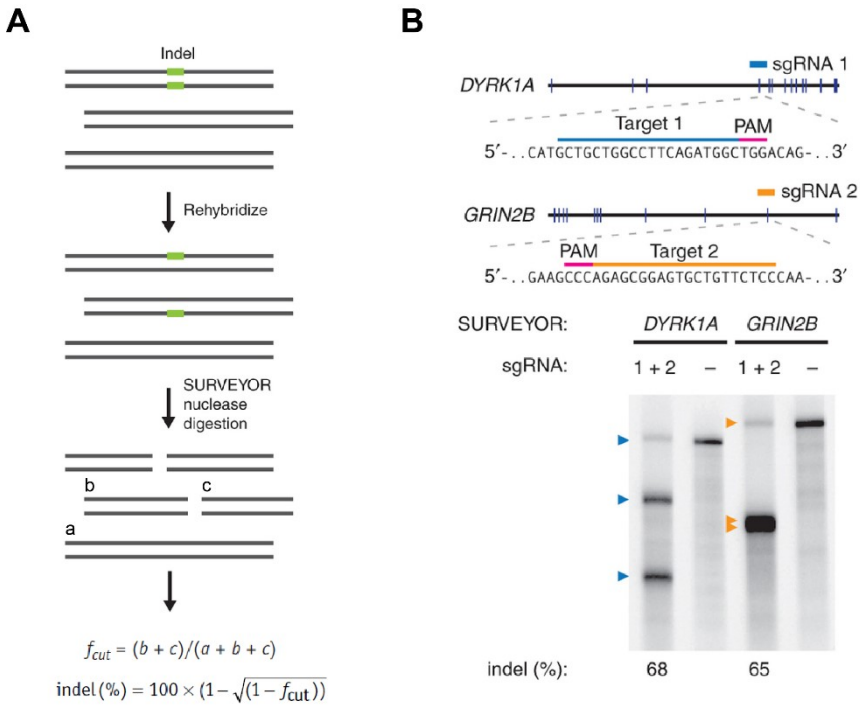


Figure 5.5 Schematic of SURVEYOR nuclease assays for genotyping clonal lines. (reproduced from [271]). (A) Schematic of the SURVEYOR nuclease assay. First, genomic DNA from the heterogeneous population of Cas9-targeted cells is isolated and amplified by PCR producing amplicons around 200-800 bp surrounding the Cas9 target site. PCR products are then reannealed to generate heteroduplexes. The reannealed heteroduplexes are cleaved by SURVEYOR nuclease, whereas homoduplexes are left intact. Cas9-mediated cleavage efficiency (percentage indel) is calculated on the basis of the fraction of cleaved DNA, as determined by the integrated intensity of gel bands. (b) An example of SURVEYOR nuclease assay. Two sgRNAs (orange and dark blue bars) are used to target the human *GRIN2B* and *DYRK1A* loci. SURVEYOR gel shows modification at both loci in transfected cells. Coloured arrowheads indicate expected fragment sizes for each locus.

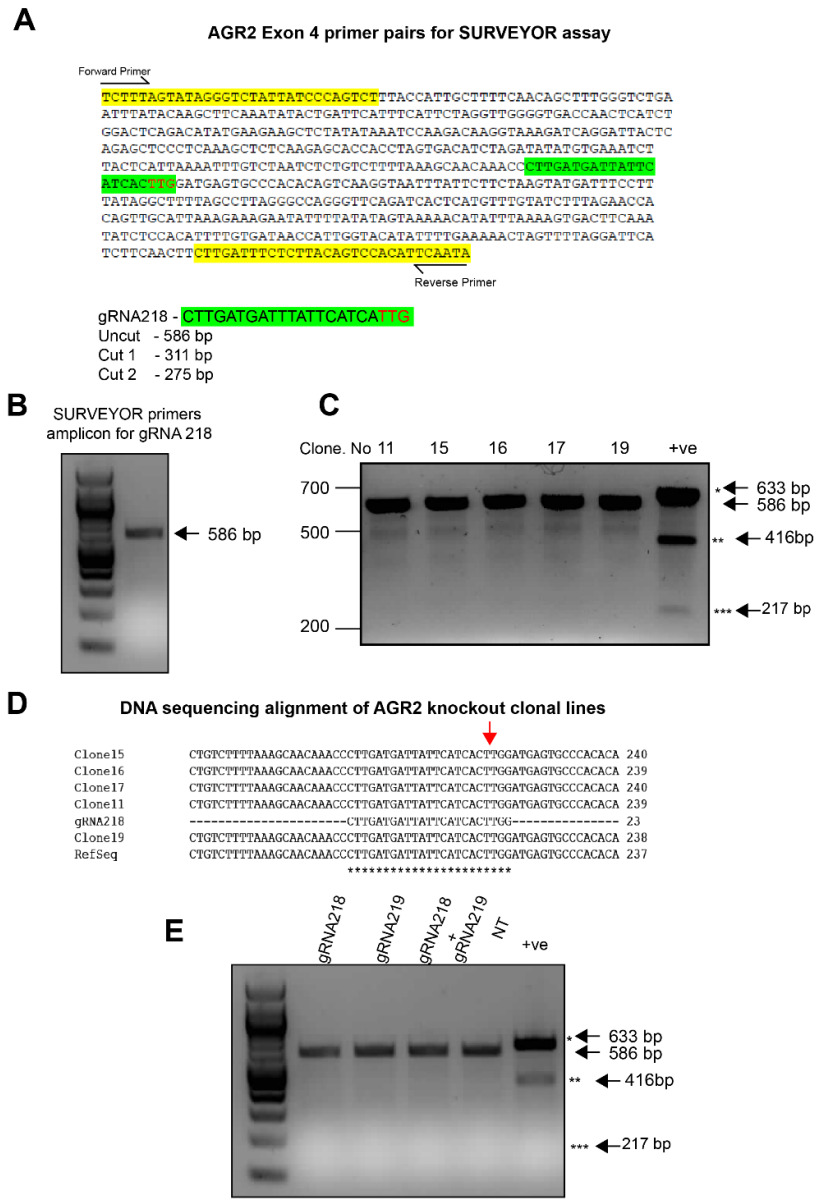


Figure 5.6 SURVEYOR nuclease assay of AGR2 knockout clonal line candidates. (A) The design of SURVEYOR primers (yellow) for AGR2 knockout within AGR2 exon 4 where the gRNA218 is located (green) that will amplify 586 bp long. SURVEYOR nuclease will recognise the indels occurring near Cas-9 cleavage site (3' bp PAM site) which will produce two more fragments; ~311 bp and ~275 bp fragments. (B) Examination of SURVEYOR primers for gRNA218 from non-transfected MCF-7 cells, which produce a single band at the expected size 586bp. (C) SURVEYOR nuclease assays of the clonal line candidates for AGR2 knockout. Positive control was run from the SURVEYOR assay kit which detects SNP in the heteroduplex suggesting the nuclease is functional and sensitive. (D) The selected clonal lines were subjected to DNA sequencing using forward SURVEYOR primer, and the sequences were aligned to see any cleavage site around the 3' upstream PAM (red arrow) (E) MCF-7 cells were transfected with gRNA218 or gRNA219 or in combination of both and incubated for 72 hours before the cells are harvested for SURVEYOR assay. These transfected cells contain a pool of both edited and unedited genome which can be assayed by SURVEYOR nuclease.

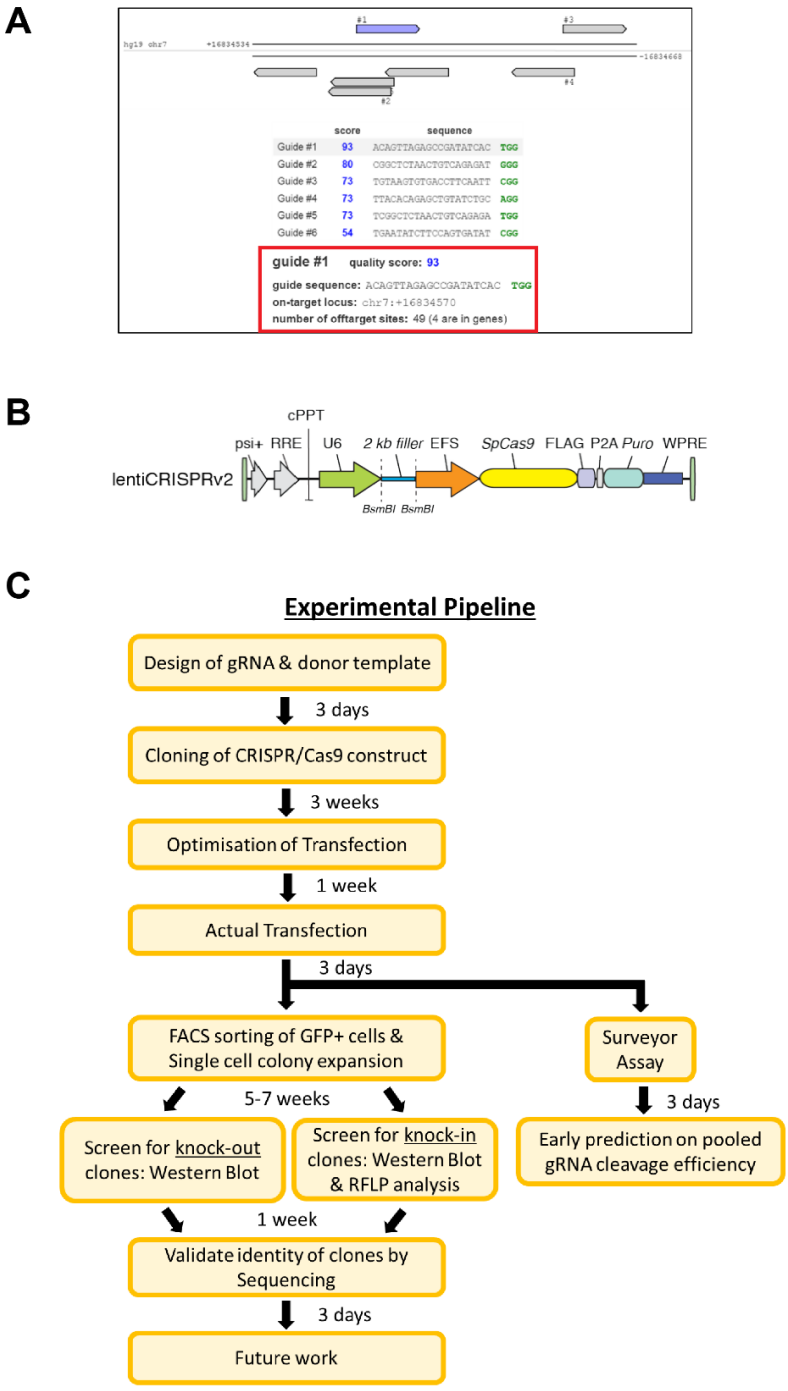


Figure 5.7 New gRNA design for AGR2 knockout and experimental pipeline (Courtesy of Lee Puay Ling). (A) New gRNA design using Feng Zhang lab gRNA design tool at Massachusetts Institute of Technology (MIT) (http://crispr.mit.edu/) targeting exon 7 of AGR2 genomic DNA. The gRNA has a high-quality score and few off-targets. (B) The gRNA was cloned in-house into pLentiCRISPRv2 using *BsmBI*. The vector contains a U6 promoter, expressing wild-type Cas9 which is under the control of CMV promoter. This plasmid also contains a 2A puromycin resistance gene to allow screening and selection of transfected cells. (C) New experimental pipeline for knocking out AGR2.

5.2.1.6. Generation of stable AGR2 expressing cells.

Due to unsuccessful attempts to knockout AGR2 using CRISPR/Cas9 technology, we next aim to use gain-of-function analysis to understand AGR2 function in cells. This can be done by overexpressing AGR2 in cells which in turn can reprogram the cells that can involve inhibition or activation of a protein, a complex, or a pathway by different molecular mechanisms. The changes in global proteome triggered by AGR2 overexpression can be monitored by the use of quantitative proteomics and the outputs can be used to create interaction network.

Transfection into a cell can be in either two forms; transient or stable. In transient transfection, the plasmid DNA is inserted into the nucleus of the cell but does not get integrated into the chromosome. The foreign gene in the nucleus will be temporarily be expressed. Since the gene is not integrated into the genome, the transfected gene does not pass to future generations of the cell. Stable transfection, on the other hand, does incorporate the plasmid DNA into the genome and gets passed on to future generations of the cell. For stable transfection to take place, the plasmid DNA must carry selectable marker (e.g antibiotic resistance) on the DNA construct or on a separate vector that is co-transfected into the cell. Then appropriate selective pressure is applied to the cells after a short recovery period.

We initially sought to establish a stable cell line expressing AGR2 and its client protein EpCAM in two different cell lines; MCF-7 (AGR+/EpCAM+) and FLO-1 (AGR2-/EpCAM-) cells. These stable cell lines can be then used to manipulate AGR2 and EpCAM interaction and pathways, especially in null FLO-1 cells. For example, by making stable EpCAM expressing cells in FLO-1 (AGR2-/EpCAM-), we could ask how the introduction of AGR2 ectopically will impact the expression and localisation of membrane-bound EpCAM? Will it be the same as the transient reconstitution of AGR2 and EpCAM were previously seen? What are the effects on growth, adhesion and metastasis?

We chose EpCAM-GFP and AGR2-CHERRY constructs for making stable cell lines in MCF- and FLO-1, as both vectors contain a selectable neomycin antibiotic

resistance gene. These constructs also express fluorescently-tagged proteins that can be visualised directly by fluorescent microscope, so any manipulation of the cells can be assayed quantitatively. MCF-7 and FLO-1 cells were transfected with each of the following; i) GFP-expressing vector; ii) mCHERRY-expressing vector; iii) EpCAM-GFP or iv) AGR2-CHERRY and incubated for 72 hours before selection process by adding G418 (an analogue of neomycin). Five colonies of each transfected DNA from each cell line were then transferred into new wells and expanded. Analysis of the stable cell lines in both FLO-1 and MCF-7 expressing EpCAM-GFP and AGR2-CHERRY by fluorescence microscopy demonstrated that most cells did not express the fusion proteins except for cells transfected with GFP and mCHERRY-expressing vectors. There were few colonies that express the fusion proteins but the number of cells expressing the fusion proteins were relatively low (~<5% cells from total cells). This suggests that creating the stable cell line expressing either EpCAM-GFP or AGR2-CHERRY were not expressed well or perhaps both proteins were secreted resulted in a low percentage of cells having the fluorescent proteins.

5.2.1.7. Real-time visualisation and quantification of cells expressing EpCAM-GFP and AGR2-GFP

We then investigated the cells expressing EpCAM-GFP or AGR-CHERRY in real-time quantitative live-cell analysis from the point of the transfection until 3 days post-transfection. This analysis can help us to determine the number of cells expressing the fusion proteins over time and explain why a generation of the stable cell line expressing both proteins was failed. For this analysis, we used IncuCyte® ZOOM System (Essen Bioscience) which can monitor and quantify cell behaviour over time and also supports HD phase-contrast, green fluorescence and red fluorescence automated imaging modes. We used AGR2-GFP instead of AGR2-CHERRY so that the image acquisition (green fluorescence) can be done simultaneously with EpCAM-GFP, and these were transfected into cells using a 96-well plate.

MCF-7 and FLO-1 cells were transfected with each of the following; i) GFP-expressing vector; ii) EpCAM-GFP, and iii) AGR2-GFP and incubated for 72 hours. The cells were transfected with the same amount of DNA but with three different cell

densities, each in triplicates; 4000, 8000 and 12000 cells per well of 96-well plate (Figure 5.8). Image acquisition of the cells (HD phase contrast and green fluorescence) were taken every 3 hours for a total of 72 hours. After 72 hours, the images were analysed as a function of the percentage of green fluorescence confluency (GFP) of the transfected cells (Number of GFP cells/Total number of cells (phase-contrast) over time. In FLO-1 cells, the number of cells with AGR2-GFP expression was very low or non-existent in all the three densities (Figure 5.8A-C). For EpCAM-GFP, the cells with green fluorescence increased after about 24 hours' post-transfection but then gradually decreased over the time period (Figure 5.8A-C). Although there was an increased number of cells with EpCAM-GFP, the highest percentage of GFP expression was approximately 5% (Figure 5.8C). In contrast, the GFP-expressing vector only, the GFP expression increased over time in all cases and reached as high as 20% (Figure 5.8A-C). Almost the same observations were seen with MCF-7 cells. Cells with AGR2-GFP expression showed a slight increase of GFP fluorescence after 20 hours (~2%) but then dropped to nearly zero (Figure 5.8D-F). The EpCAM-GFP expression was maximum after 24 hours' post-transfection and then plateaued for the rest of incubation period (Figure 5.8D-F). MCF-7 cells transfected with the GFP-expressing vector showed similar observation in FLO-1 cells, where the expression increased over time (Figure 5.8D-F).

This data suggest that AGR2-GFP expression is not well maintained in either AGR2-negative cells (FLO-1) or AGR2-positive cells(MCF-7). This may be because the AGR2-GFP is secreted as described before. AGR2 has also been shown to have adhesive property where it can be coated to the plastic substratum, and the coated AGR2 protein was able to promote cell adhesion of rat mammary tumour cells [61]. This finding leads us to another hypothesis where AGR2 expression is 'sticky' to the cells which can prevent cells to proliferate. The low expression of AGR2-GFP in our real-time analysis explains why we could not isolate stable cell clone expressing AGR2-GFP. In the case of EpCAM-GFP, the expression is not well-maintained in FLO-1 which is EpCAM negative-cells, probably due to no authentic EpCAM synthesis pathway and overexpressing the protein could be toxic to cell over a period of time (expression decrease after 20 hours Figure 5.8A-C). In contrast in EpCAM-

positive cells, MCF-7, EpCAM-GFP expression is maintained after 20 hours, suggesting that there is a feedback mechanism to maintain the steady-state level of EpCAM-GFP expression. This is not the case with GFP-vector control only where its expression is comparatively high in both FLO-1 and MCF-7 cells, implying that that AGR2 and EpCAM have different growth characteristics.

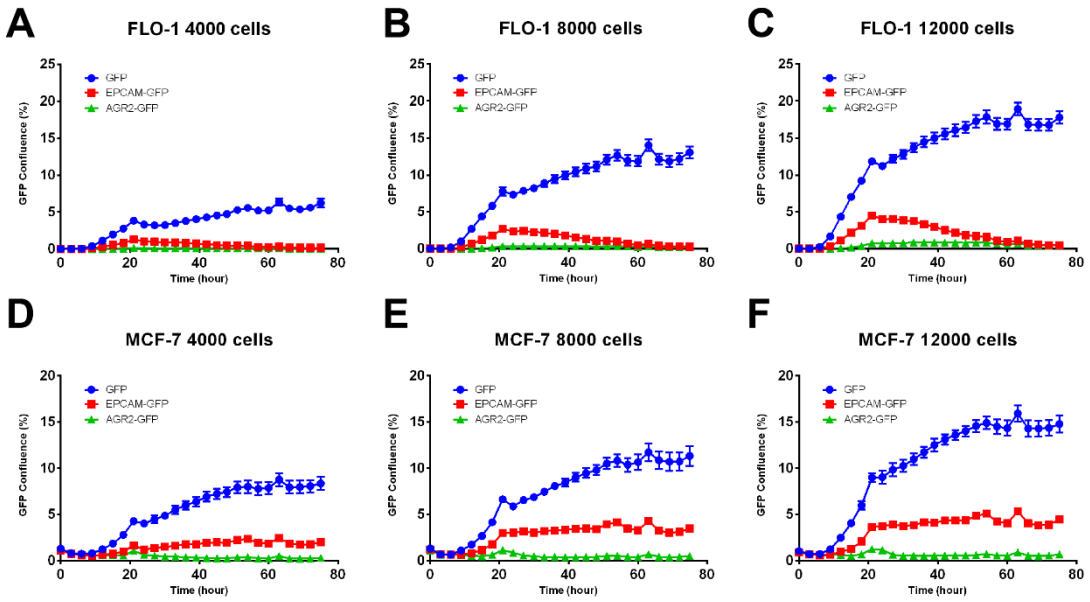


Figure 5.8 Real-time visualisation and quantification of cells expressing EpCAM-GFP and AGR2-GFP. MCF-7 and FLO-1 cells were transfected with either: i) GFP-expressing vector; ii) EpCAM-GFP, or iii) AGR2-GFP and incubated for 72 hours in incubator equipped with IncuCyte® ZOOM System. The cells were transfected with the same amount of DNA but with three different cell densities, each in triplicates; (A and D) 4000, (B and E) 8000 and (C and F) 12000 cells per well of 96-wellplate. Image acquisition of the cells (HD phase contrast and green fluorescence) were taken every 3 hours for a total of 72 hours. After 72 hours, the images were analysed as a function of the percentage of green fluorescence confluency (GFP) of the transfected cells (Number of GFP cells/Total number of cells (phase-contrast) over time.The experiments were performed in three independent experiments in triplicates and data presented here are representative of one experiment. Error bars represent SEM from three biological replicates from one experiment.

5.2.3. Live-cell fluorescence imaging of FLO-1 transfected with AGR2-CHERRY

We also evaluated the expression of AGR2-CHERRY in FLO-1 cells using live-cell fluorescence microscopy to further assess AGR2 expression and to monitor its localisation and movement in the AGR2-null cell. FLO-1 cells coated on coverslips were transfected with AGR2-CHERRY construct and incubated for 24 hours. The coverslips containing the transfected cells were then subjected to live-cell imaging using epifluorescence microscopy equipped with incubation chambers set at physiological conditions. The results demonstrated that AGR2-CHERRY showed cytoplasmic localisation and a subset of the population showed puncta staining (Figure 5.9, left panel). When AGR2-CHERRY was subjected to live-recordings, the puncta moved rapidly in the cell suggesting that AGR2-CHERRY localised in secretory vesicles. The movement of the vesicles corresponding to AGR2-CHERRY were tracked and colour-coded as a function of time (0-50ms) (Figure 5.9, right panel). The image analysis proved that these vesicles moved at a high-rate throughout the cell. These are unusual features of an ER-resident protein (i.e localisation of ER-resident protein like calreticulin or Sec61β generally show tubular or flat sheet structure) and further suggest that AGR2 can escape the ER or 'bud off' the ER for its secretion or perhaps AGR2 involve in chaperoning proteins destined for secretory pathway.

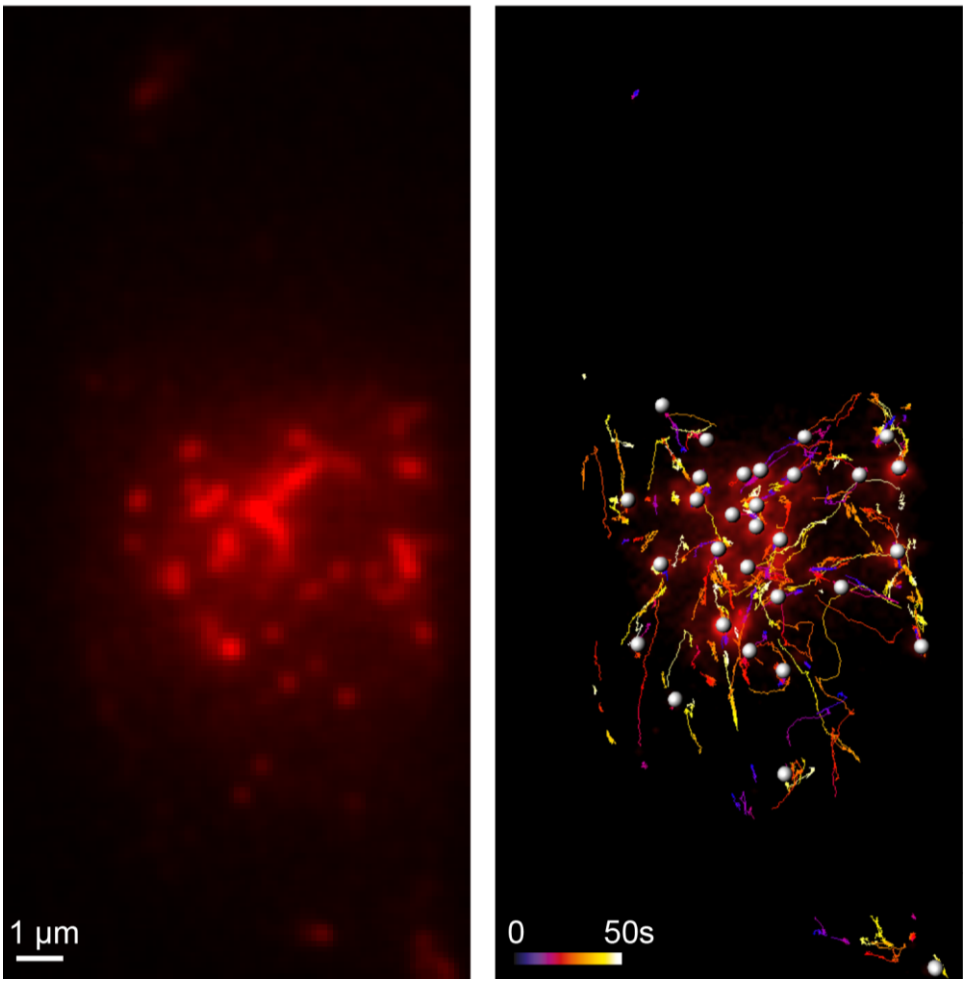


Figure 5.9 Live-cell fluorescence imaging of AGR2-CHERRY in FLO-1 cells. FLO-1 cells grown in 25mm coverslip in 6-well plate (SLS Cat No. MIC3350) were transfected with AGR2-CHERRY construct (2µg) and incubated for 24 hours. Cells on coverslips were transferred onto coverslip chamber onto microscope incubation chamber set at 37^0C and 5% CO_2. Images and live-recordings (100x time lapse) were captured in widefield using Olympus IX81 epiflourescence microscopy with 63X objective. Vesicles tracking and analysis was performed using Imaris by Dr Alison Dun, Heriot-Watt University.

5.2.4. Reconstitution of AGR2 by transient transfection

5.2.4.1. Expression of AGR2 by electroporation

After having unsuccessful attempts to generate isogenic cell line by knocking out AGR2 by CRISPR/Cas9 tool and to generate stable cell line overexpressing AGR2, we decided to reconstitute AGR2 into AGR2-null cells (FLO-1) using transient transfection and process for mass-spectrometry. Live-imaging of AGR2-CHERRY in FLO-1 cells suggests the cells can express AGR2-CHERRY and has intact ER in which AGR2 can 'bud off' for its dedicated function. The reconstitution of AGR2 in FLO-1 results in transient overexpression that can be used to study AGR2 gain-of-function. We chose to deliver the DNA constructs using mammalian electroporation system. This is to ensure that the cells are free from any lipids that associated with most transfection reagents like Attractene and Lipofectamine 2000 which are lipid-based that could complicate mass spectrometry quantitation. Also, electroporation offers higher transfection efficiency compared to traditional lipid-based transfection reagents which we have confirmed in our lab. We compared the transfection efficiency of GFP-expressing vector (pEGFP-N1 and pmaxGFP) using Attractene and electroporation system (Nucleofector® Device, Lonza) after 24 hours incubation and counted the GFP- positive cells using a flow cytometry. The transfection efficiency was about 4-fold higher using electroporation (data not shown).

We opted to use untagged-AGR2 and not the fluorescently-tagged AGR2 to make sure that AGR2 is responsible for reprogramming the cells when we screen for any proteomic changes. We have cloned AGR2 in two different vectors with different promoters as described in Chapter 4; CMV (strong promoter) and RSV (weak promoter). Both AGR2 constructs expressed well and as expected in cells where the AGR2 in CMV vector express higher than its RSV counterpart. Since AGR2 expression is poor in FLO-1 cells and the fact that AGR2 can be 'adhesive', we decided to choose the low expression promoter RSV so that overexpression of AGR2 is not toxic to the cell. The introduction of low expressing AGR2 into cells also ensures true representative of AGR2-reprogrammed cells, as any dramatic changes in cells can lead to induction of unexpected pathways that could complicate our proteomics screen

analysis. We also showed that AGR2 structural variant with the more classic ER retention motif (C-terminal KTEL mutated to KDEL) can also stimulate AGR2 client protein EpCAM production and delivery to the plasma membrane (Chapter 4) and included in our proteomic screen.

FLO-1 cells were transfected in biological triplicates with each of the following constructs in a 6-well plate; i) RSV empty vector (pSF-RSV) ii) wt-AGR2 (pAM1L) iii) AGR2-KDEL (pAM2L). The cells were incubated for 72 hours. The cells were then harvested as described in Materials and Methods and lysed using urea lysis buffer. One of the triplicates were subjected to western blot analysis to confirm the expression, while the other two were prepared for mass spectrometry and were only performed once. Western blot analysis confirmed the expression of wt-AGR2 and AGR2-KDEL in FLO-1 cells, while the empty vector RSV (negative control) showed no expression of AGR2 (Figure 5.10). The AGR2-KDEL showed higher expression than the wt-AGR2 which we previously observed in Chapter 4. This is also due to the fact that the wt-AGR2 can be secreted into media hence the lower expression, while AGR2-KDEL with enhanced ER localisation is retained intracellularly, therefore higher expression.

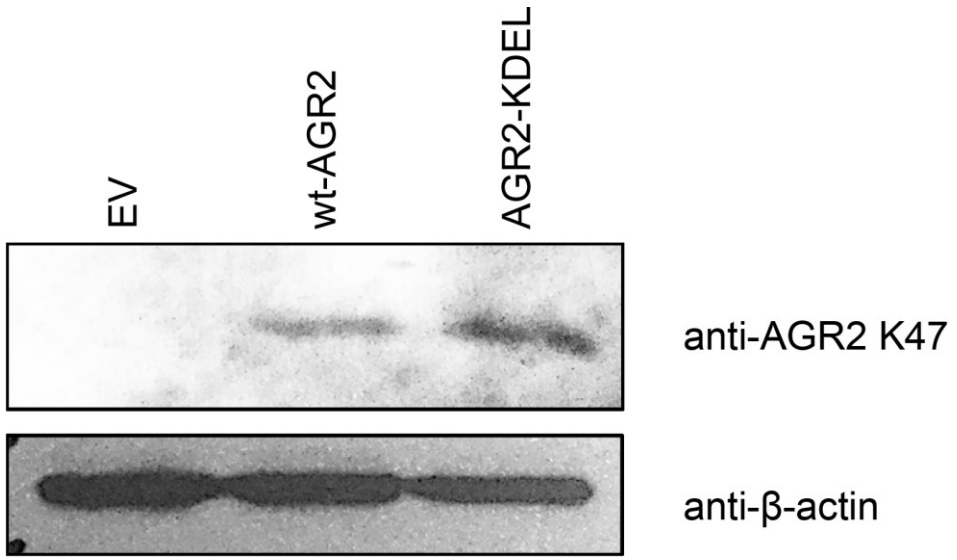


Figure 5.10 Western blot analysis of FLO-1 cells expressing wt-AGR2, AGR2-KDEL, or RSV empty vector. FLO-1 cells were transfected with 2 μg of either: i) RSV empty vector (EV) ii) wt-AGR2-RSV or iii) AGR2-KDEL-RSV in a 6-well plate in triplicates using mammalian electroporation system. The cells were incubated for 72 hours and one of the triplicates were harvested using urea lysis buffer and subjected to western blot as showed here using anti-AGR2 K47 and anti-β-actin as a loading control.

5.2.5. The application of quantitative proteomics using tandem mass tag mass spectrometry

Tandem mass tag (TMT) mass spectrometry is a quantitative proteomics approach to accurately determine peptides and proteins in a complex sample (Figure 5.11). During sample preparation, fractionated peptides are labelled with TMT tag. Each tag comprises a sensitization group and a mass differentiated group that together comprise the TMT fragment which is actually detected [272]. TMT tags consist of an amine reactive, N-hydroxysuccinimide ester (NHS-ester), balancing carbonyl linker and reporter ion components. The tags also comprise a reactive functionality (R) to enable the tag to be coupled to any peptide which forms amide linkages with the label peptides via N-termini or lysine residues. TMT mass spectrometry allows multiplexing, where multiple tags with different reporter ions can be used simultaneously for each different samples and all the samples are mixed prior to fractionation and analysed using tandem mass spectrometry (MS). TMT-tagged peptides are chemically identical and have the same total mass and chromatographic separation and therefore the liquid chromatography (LC) retention time and mass/charge (m/z) separation of each labelled peptides do not split into different peaks during the MS1 scan [273]. TMT strategy relies on dissociation of the TMT fragment by a number of methods of dissociation. One such method involves colliding the ions with a stream of inert gas, which is known as collision-induced dissociation (CID) or higher energy collision dissociation (HCD) to produce ionised labelled peptide in the MS2 mode (Figure 5.11). The peptide fragments are detected by generating mass spectra in the usual manner fragment. The reporter ions are detected and as peaks at a predefined m/z. For a TMTsixplex experiment, the reporters are detected at 126.1,127.1, 128.1, 129.1, 130.1 and 131.1 m/z [272]. TMT offers better signal-to-noise ratios and permits untagged material to be discounted which significantly improving data quality.

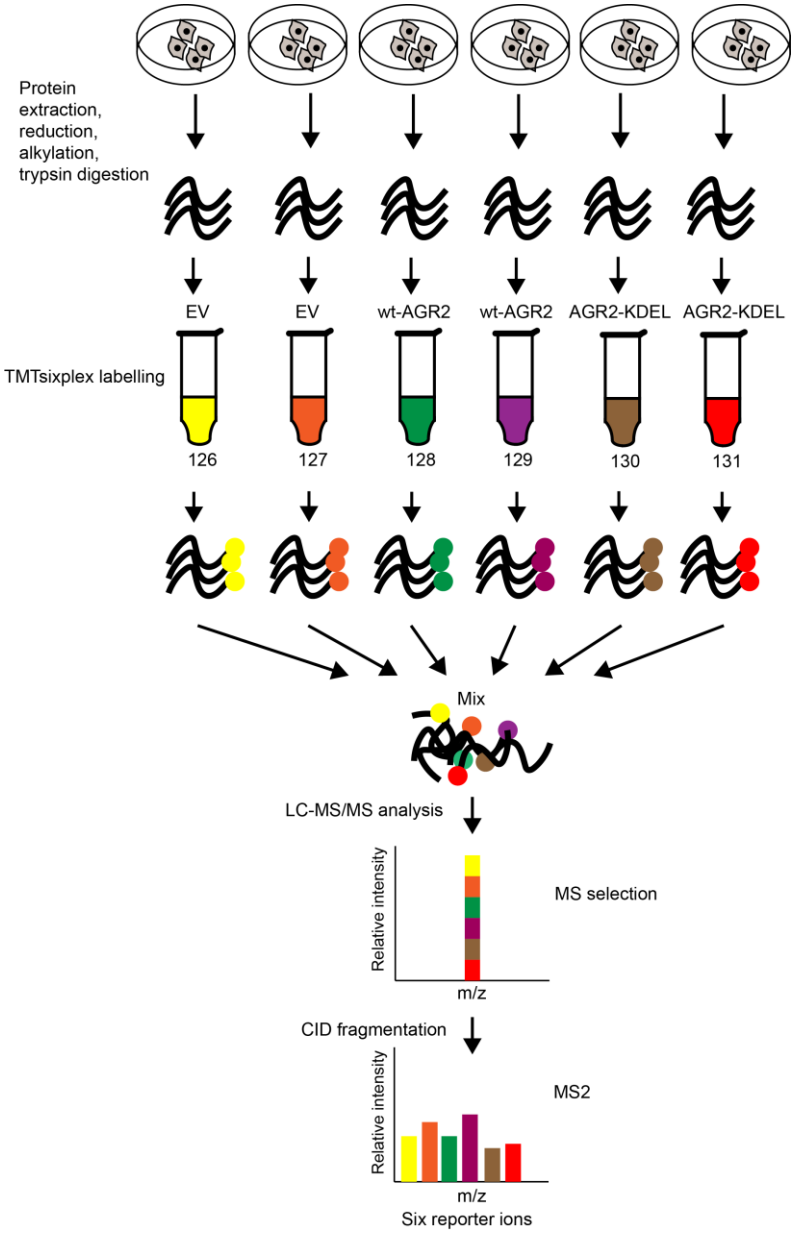


Figure 5.11 Flowchart of experimental design for TMT mass spectrometry. FLO-1 cells were transfected either: i) RSV empty vector (EV) ii) wt-AGR2-RSV or iii) AGR2-KDEL-RSV in a 6-well plate in biological triplicates using mammalian electroporation system. The cells were incubated for 72 hours and two of the triplicates were analysed by MS and MS was only performed once. Cells were lysed, reduced, alkylated and trypsin digested using FASP protocol. Peptides were labelled using TMTsixplex isobaric tags as illustrated. Labelled peptides were mixed and analysed using Orbitrap LC-MS/MS before data analysis. Labelled peptide ions are indistinguishable in the full MS. Quantification is based on low *m/z* reporter fragment relative ion intensities after a target ion is isolated and fragmented (CID/HCD) in MS2 experiment.

5.2.5.1. TMT quantitation of proteome of cells with incorporated AGR2

The TMT methodology allows a degree of multiplexing of the experiment design, to incorporate relevant controls for the analysis of the function of the AGR2 gene of interest. We employed TMTsixplex™ (Thermo Fischer Scientific) technique, allowing quantitation of three different samples/conditions in duplicates to be compared concurrently which are; cells with the RSV empty vector, cells with wt-AGR2, and cells with AGR2-KDEL. Any modification from the cell with empty vector suggests a potential role of the introduced gene in the reprogramming of the cell. This also allows us to see if there are any overlaps of dysregulated proteins from proteomics screen between the two variants, where wt-AGR2 represent ER-localised and secreted form while AGR2-KDEL represent enhanced ER-localised form.

Proteins extracted from transfected FLO-1 cells in biological duplicates are reduced, alkylated and digested as described in Materials and Methods. Samples were labelled with the TMT Reagents as illustrated in Figure 5.11. Each biological duplicate was labelled with TMT labels; EV with 126 and 127; wt-AGR2 with 128 and 19; and AGR2-KDEL with 130 and 131. Labelled peptides were then mixed following by sample fractionation and clean up. Samples were then analysed by high-resolution Orbitrap LC-MS/MS before data analysis to identify peptides and to quantify reporter ion relative abundance, thanks to Dr Lenka Hernychova. The total number of proteins identified in the spectral files were 1224 (Appendix 5). The quantitation of peptides was further analysed and the relative abundance of the peptides was measured by comparing the peptide pairs labelled with different TMT tag (e.g 128/126 for wt-AGR2/Control) followed by statistical evaluation on the peptide ratios to infer protein levels. The fold changes of quantified proteins in biological duplicates were normalised by calculating the geometric mean of the of the fold changes. The fold changes of all quantified proteins for each AGR2 expressing cells over control were plotted against the TMT scores (Figure 5.12A-B). The majority of protein expression levels quantified across all conditions were not dramatically altered from the control AGR2-null cell line. However, we noticed that for wt-AGR2 expressing cells the global proteins seem to be clustered to the left quadrant suggesting most proteins were

downregulated. On the other hand, the more enhanced ER-localised AGR2, AGR2-KDEL expressing cells showed most proteins clustered to the right quadrant suggesting that most proteins were upregulated. This suggests that the two AGR2 variants have specific distinct signalling pathways consistent with the theory that both proteins have different localisations in the cells. We also plotted the fold changes of wt-AGR2 and AGR2-KDEL against each other so we can determine overlapping proteins between the two populations (Figure 5.12C). The plot shows clustering of proteins towards the second quadrant (upper left) suggesting that most proteins were upregulated in AGR2-KDEL but the same set of proteins were downregulated in wt-AGR2 cells. Only a few proteins were found to be upregulated in both conditions and the key proteins were highlighted in green (Figure 5.12C, first quadrant) such as AGR2 itself, AK3, MRPS28, RAB6A and DCD. Two proteins were found to be the dominant downregulated outliers and these are KRT17 and PAX highlighted in blue (Figure 5.12C, third quadrant).

The database mining of proteins containing AGR2 docking motif shows enrichment in transmembrane proteins (Chapter 3), suggesting that AGR2 involve in the maturation of this class of protein, at least in the case of transmembrane EpCAM. Therefore, we expect that introduction of AGR2 will yield upregulation of this class of proteins. However, when we ranked the differentially expressed proteins, the top 10 most upregulated and downregulated proteins in wt-AGR2 or AGR2-KDEL expressing cells listed were not enriched in transmembrane proteins. This might be that the transmembrane proteins were not properly extracted during sample preparation due to highly hydrophobic and water-insoluble nature of membrane proteins the fact that they are present at low abundance and the absence of polar side chains in amino acid residues [274]. The use of shotgun proteomics or enrichment of glycoproteins during sample preparation might be a better idea to resolve identification of membrane protein landscape [275, 276].

We further analysed the top 40 most upregulated proteins in wt-AGR2 and AGR2-KDEL expressing cells and subjected to functional enrichment tools using FunRich [212]. When the top 40 proteins upregulated in wt-AGR2 expressing cells

were scored based on subcellular components, the majority of the proteins have localisation in the exosomes and lysosomes. The AGR2-KDEL expressing cells also showed the same enrichment and additionally displayed protein localisations in the cytoplasm.

We also looked at the overlapping proteins of the top 40 upregulated genes in both cases of AGR2-expressing cells. There are 15 overlapping proteins suggesting potential AGR2 client proteins. These 15 proteins mainly enriched in lysosomal proteins such as SEC24 family member C (SEC24C) and superoxide dismutase 1 (SOD1). However, these proteins, are not validated in this study. We chose dermcidin (DCD) protein as a potential AGR2 binding protein. DCD is upregulated in both wt-AGR2 and AGR2-KDEL expressing cells and the fact that DCD is also found in another proteomic screen in the lab where it was identified in label-free SWATH (sequential window acquisition of all theoretical fragment ion spectra**)** mass spectrometry analysis of ELMO1 pull-down assay from FLO-1 and OE19 cells (Ali Alsaadi Thesis 2016, and personal communication). ELMO1 was shown to form protein-protein interaction complex with AGR2 as well as DCD in cells using in situ PLA, suggesting that AGR2 may potentially bind to DCD.

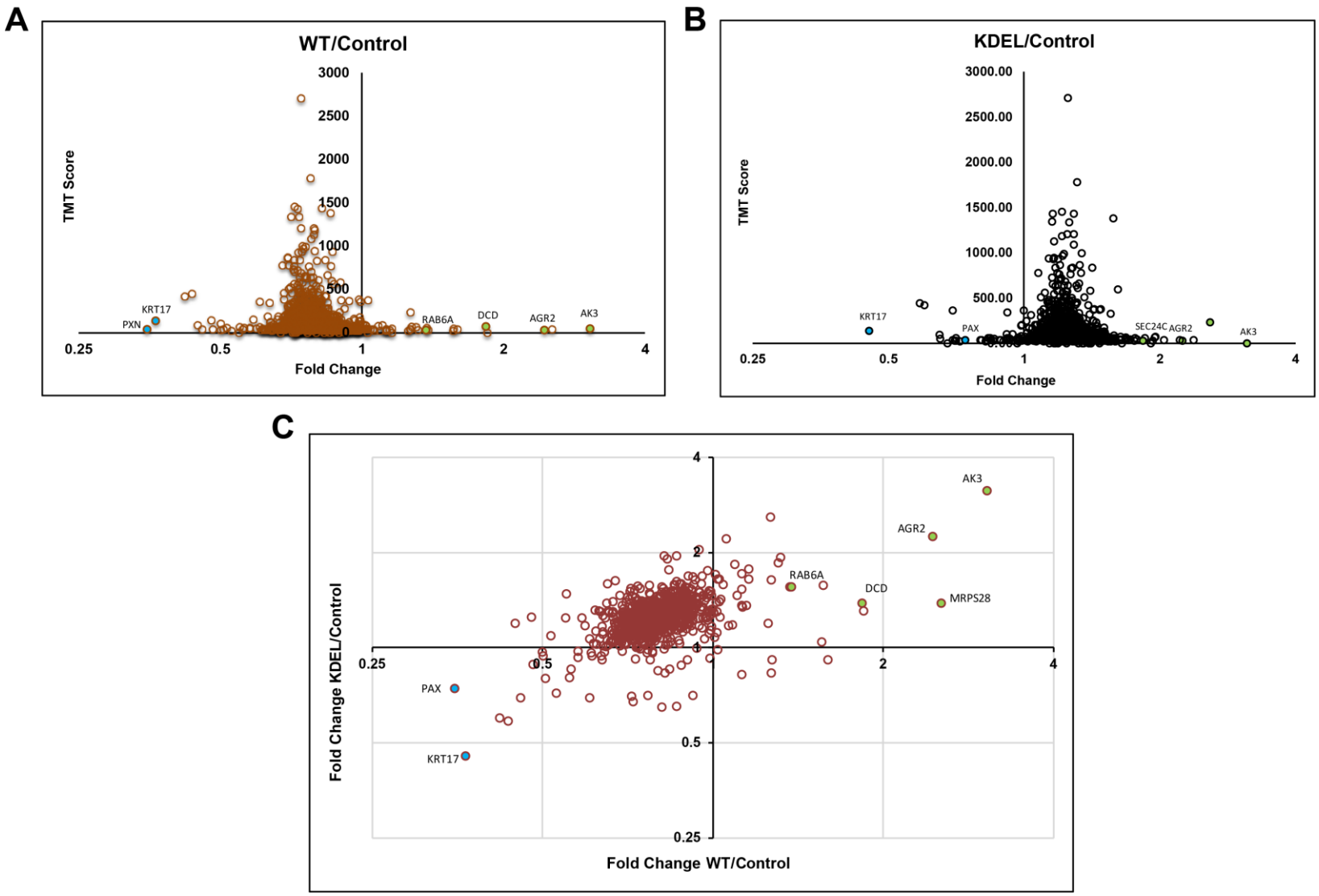

A
WT/Control
TMT Score
3000
2500
2000
1500
1000
500
0
KRT17
PXN
RAB6A
DCD
AGR2
AK3
0.25
0.5
1
2
4
Fold Change
B
KDEL/Control
TMT Score
3000.00
2500.00
2000.00
1500.00
1000.00
500.00
KRT17
PAX
SEC24C
AGR2
AK3
0.25
0.5
1
2
4
Fold Change
C
Fold Change KDEL/Control
4
2
1
0.5
0.25
AK3
AGR2
RAB6A
DCD
MRPS28
PAX
KRT17
0.25
0.5
2
4
Fold Change WT/Control

Figure 5.12 Quantitative mass spectral data of wt-AGR2 and AGR2-KDEL expression subtly reprograms the proteome of the cell. (A) Distribution of fold changes of all proteins quantified in wt-AGR2 compared to EV control FLO-1 cells that are plotted as a function of TMT scores. Most proteins showed downregulation of the proteome as they are clustered to the left quadrant, while few proteins were upregulated showing in the right quadrant. (B) Distribution of fold changes of all proteins quantified in AGR2/KDEL compared to EV control FLO-1 cells that are plotted as a function of TMT scores. Most proteins showed upregulation of the proteome as they are clustered to the right quadrant, while few proteins were upregulated as plotted in the left quadrant. (C) Distribution of fold changes of all quantified proteins in AGR2-KDEL/Control on the y-axis and wt-AGR2/Control on the x-axis plotting against each other, indicating common genes upregulated (quadrant 1, top right) or downregulated (quadrant 3, bottom left) in both conditions. In all plots, key upregulated proteins were annotated and marked with green colour for upregulated proteins while downregulated proteins with blue colour. All identified proteins were summarised in Appendix 5.

A 10 most upregulated proteins: AGR2 WT

Accession	Description	Normalized fold change value	Score	Log2 fold change
Q9UIJ7	GTP:AMP phosphotransferase AK3, mitochondrial OS=Homo sapiens GN=AK3 PE=1 SV=4 - [KAD3_HUMAN]	3.0582	0	1.6127
Q9Y2Q9	28S ribosomal protein S28, mitochondrial OS=Homo sapiens GN=MRPS28 PE=1 SV=1 - [RT28_HUMAN]	2.5400	31.8	1.3448
O95994	Anterior gradient protein 2 homolog OS=Homo sapiens GN=AGR2 PE=1 SV=1 - [AGR2_HUMAN]	2.4474	27.69	1.2912
Q5H9R7	Serine/threonine-protein phosphatase 6 regulatory subunit 3 OS=Homo sapiens GN=PPP6R3 PE=1 SV=2 - [PP6R3_HUMAN]	1.8466	23.79	0.8849
P81605	Dermcidin OS=Homo sapiens GN=DCD PE=1 SV=2 - [DCD_HUMAN]	1.8401	70.43095	0.8798
P61163	Alpha-centractin OS=Homo sapiens GN=ACTR1A PE=1 SV=1 - [ACTZ_HUMAN]	1.5968	33.77	0.6752
P43246	DNA mismatch repair protein Msh2 OS=Homo sapiens GN=MSH2 PE=1 SV=1 - [MSH2_HUMAN]	1.5679	57.97	0.6488
P18583	Protein SON OS=Homo sapiens GN=SON PE=1 SV=4 - [SON_HUMAN]	1.5599	51.1	0.6414
P20340	Ras-related protein Rab-6A OS=Homo sapiens GN=RAB6A PE=1 SV=3 - [RAB6A_HUMAN]	1.3815	33.5	0.4662
Q99627	COP9 signalosome complex subunit 8 OS=Homo sapiens GN=COPS8 PE=1 SV=1 - [CSN8_HUMAN]	1.3746	20.35	0.4590

B 10 most downregulated proteins: AGR2 WT

Accession	Description	Normalized fold change value	Score	Log2 fold change
P49023	Paxillin OS=Homo sapiens GN=PXN PE=1 SV=3 - [PAXI_HUMAN]	0.3504	33.2400	-1.5130
Q04695	Keratin, type I cytoskeletal 17 OS=Homo sapiens GN=KRT17 PE=1 SV=2 - [K1C17_HUMAN]	0.3657	135.6991	-1.4513
P08779	Keratin, type I cytoskeletal 16 OS=Homo sapiens GN=KRT16 PE=1 SV=4 - [K1C16_HUMAN]	0.4204	418.2377	-1.2503
P02538	Keratin, type II cytoskeletal 6A OS=Homo sapiens GN=KRT6A PE=1 SV=3 - [K2C6A_HUMAN]	0.4348	445.3933	-1.2016
P02765	Alpha-2-HS-glycoprotein OS=Homo sapiens GN=AHSG PE=1 SV=1 - [FETUA_HUMAN]	0.4471	85.0233	-1.1614
Q92979	Ribosomal RNA small subunit methyltransferase NEP1 OS=Homo sapiens GN=EMG1 PE=1 SV=4 - [NEP1_HUMAN]	0.4579	36.6500	-1.1270
Q9BUF5	Tubulin beta-6 chain OS=Homo sapiens GN=TUBB6 PE=1 SV=1 - [TBB6_HUMAN]	0.4780	145.8718	-1.0650
P19388	DNA-directed RNA polymerases I, II, and III subunit RPABC1 OS=Homo sapiens GN=POLR2E PE=1 SV=4 - [RPAB1_HUMAN]	0.4816	39.0500	-1.0541
Q494V2	Coiled-coil domain-containing protein 37 OS=Homo sapiens GN=CCDC37 PE=2 SV=1 - [CCD37_HUMAN]	0.5000	29.0600	-0.9999
P02768	Serum albumin OS=Homo sapiens GN=ALB PE=1 SV=2 - [ALBU_HUMAN]	0.5012	101.3700	-0.9965

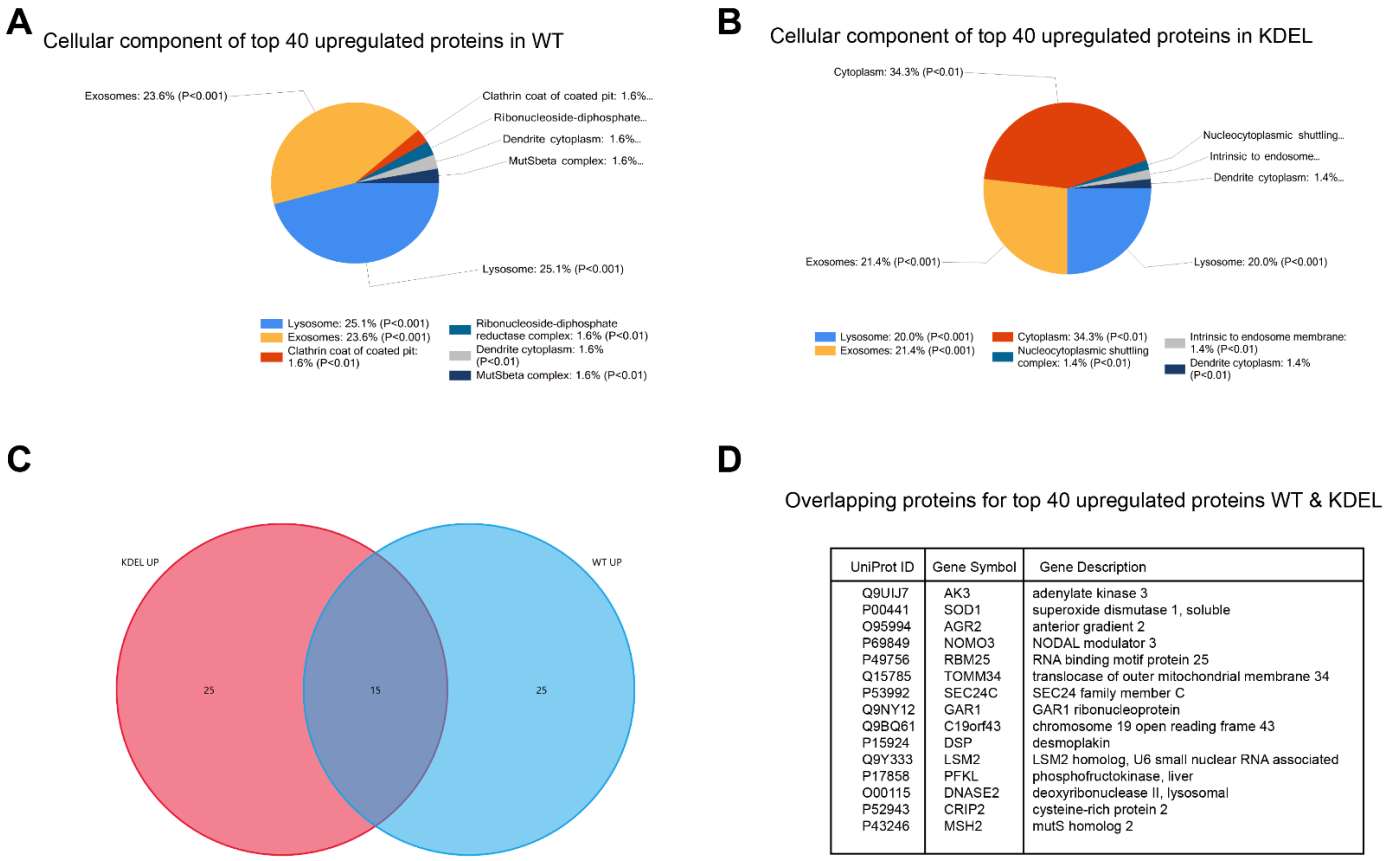
C 10 most upregulated proteins: AGR2 KDEL

Accession	Description	Normalized fold change value	Score	Log2 fold change
Q9UIJ7	GTP:AMP phosphotransferase AK3, mitochondrial OS=Homo sapiens GN=AK3 PE=1 SV=4 - [KAD3_HUMAN]	3.1287	0.0000	1.6455
P00441	Superoxide dismutase [Cu-Zn] OS=Homo sapiens GN=SOD1 PE=1 SV=2 - [SODC_HUMAN]	2.5847	232.6392	1.3700
Q13126	S-methyl-5'-thioadenosine phosphorylase OS=Homo sapiens GN=MTAP PE=1 SV=2 - [MTAP_HUMAN]	2.3794	34.1800	1.2506
O95994	Anterior gradient protein 2 homolog OS=Homo sapiens GN=AGR2 PE=1 SV=1 - [AGR2_HUMAN]	2.2441	27.6900	1.1661
P69849	Nodal modulator 3 OS=Homo sapiens GN=NOMO3 PE=2 SV=2 - [NOMO3_HUMAN]	2.2152	34.8900	1.1474
P37198	Nuclear pore glycoprotein p62 OS=Homo sapiens GN=NUP62 PE=1 SV=3 - [NUP62_HUMAN]	2.0405	39.5800	1.0289
O00764	Pyridoxal kinase OS=Homo sapiens GN=PDXK PE=1 SV=1 - [PDXK_HUMAN]	1.9531	72.9400	0.9658
P80303	Nucleobindin-2 OS=Homo sapiens GN=NUCB2 PE=1 SV=2 - [NUCB2_HUMAN]	1.9506	62.4900	0.9639
P49756	RNA-binding protein 25 OS=Homo sapiens GN=RBM25 PE=1 SV=3 - [RBM25_HUMAN]	1.9296	29.9500	0.9483
Q8N9V7	Testis- and ovary-specific PAZ domain-containing protein 1 OS=Homo sapiens GN=TOPAZ1 PE=2 SV=3 - [TOPZ1_HUMAN]	1.9105	0.0000	0.9340

D 10 most downreguletad proteins: AGR2 KDEL

Accession	Description	Normalized fold change value	Score	Log2 fold change
Q04695	Keratin, type I cytoskeletal 17 OS=Homo sapiens GN=KRT17 PE=1 SV=2 - [K1C17_HUMAN]	0.4536	135.6991	-1.1406
P02538	Keratin, type II cytoskeletal 6A OS=Homo sapiens GN=KRT6A PE=1 SV=3 - [K2C6A_HUMAN]	0.5873	445.3933	-0.7678
P08779	Keratin, type I cytoskeletal 16 OS=Homo sapiens GN=KRT16 PE=1 SV=4 - [K1C16_HUMAN]	0.6006	418.2377	-0.7355
Q9Y5L4	Mitochondrial import inner membrane translocase subunit Tim13 OS=Homo sapiens GN=TIMM13 PE=1 SV=1 - [TIM13_HUMAN]	0.6483	91.8200	-0.6253
Q5VYK3	Proteasome-associated protein ECM29 homolog OS=Homo sapiens GN=ECM29 PE=1 SV=2 - [ECM29_HUMAN]	0.6514	54.2800	-0.6184
Q96CV9	Optineurin OS=Homo sapiens GN=OPTN PE=1 SV=2 - [OPTN_HUMAN]	0.6752	0.0000	-0.5665
P13995	Bifunctional methylenetetrahydrofolate dehydrogenase/cyclohydrolase, mitochondrial OS=Homo sapiens GN=MTHFD2 PE=1 SV=2 - [MTDC_HUMAN]	0.6807	50.7600	-0.5550
Q92979	Ribosomal RNA small subunit methyltransferase NEP1 OS=Homo sapiens GN=EMG1 PE=1 SV=4 - [NEP1_HUMAN]	0.6942	36.6500	-0.5265
P02533	Keratin, type I cytoskeletal 14 OS=Homo sapiens GN=KRT14 PE=1 SV=4 - [K1C14_HUMAN]	0.6949	362.1426	-0.5251
Q96M27	Protein PRRC1 OS=Homo sapiens GN=PRRC1 PE=1 SV=1 - [PRRC1_HUMAN]	0.7007	29.3000	-0.5131

Figure 5.13 Ranked TMT expression data presenting the ten most upregulated and downregulated protein expression changes as a result of wt-AGR2 or AGR2-KDEL genes introduction into AGR2-null FLO-1 cells. The proteins are tabulated by protein ID, TMT score and fold changes.



UniProt ID	Gene Symbol	Gene Description
Q9UIJ7	AK3	adenylate kinase 3
P00441	SOD1	superoxide dismutase 1, soluble
O95994	AGR2	anterior gradient 2
P69849	NOMO3	NODAL modulator 3
P49756	RBM25	RNA binding motif protein 25
Q15785	TOMM34	translocase of outer mitochondrial membrane 34
P53992	SEC24C	SEC24 family member C
Q9NY12	GAR1	GAR1 ribonucleoprotein
Q9BQ61	C19orf43	chromosome 19 open reading frame 43
P15924	DSP	desmoplakin
Q9Y333	LSM2	LSM2 homolog, U6 small nuclear RNA associated
P17858	PFKL	phosphofructokinase, liver
O00115	DNASE2	deoxyribonuclease II, lysosomal
P52943	CRIP2	cysteine-rich protein 2
P43246	MSH2	mutS homolog 2

Figure 5.14 Functional enrichment of the top 40 most upregulated proteins in wt-AGR2 and AGR2-KDEL expressing cells using Funrich [212]. The top 40 most upregulated proteins **in** (A) wt-AGR2 expressing cells and (B) AGR2-KDEL expressing cells were scored as cellular components. (C) Venn diagram showing overlapping proteins from the top 40 upregulated proteins between wt-AGR2 and AGR2-KDEL expressing cells. (D) List of overlapping proteins in (C).

5.2.6. Dermcidin as a potential AGR2 binding partner

Dermcidin (DCD) is a secreted protein that is subsequently processed into mature peptides with distinct biological activities [277]. In normal cell physiology, DCD play a role in the antimicrobial activity, innate immunity and embryogenesis. In cancer cells, DCD has been shown to promote cancer cell survival and responsible for disease progression. For example, DCD was found to be in a high concentration in the blood of breast cancer patients and associated with poor prognosis [278]. It was also identified in a subset of invasive breast cancers and found that its expression enhanced cell proliferation and has been reported to make cells more resistant to oxidative stress compared to control cell lines [279]. The unprocessed precursor DCD is an 110 aa long, and the protein can undergo proteolytic processing forming functional domains with different functions (Figure 5.15). The functional domain generally can be divided into two groups: prodomain and antimicrobial peptide (AMP) domain [280]. The AMP domain, for example, can give rise to a functional peptide that started with 47 aa long which is constitutively expressed in sweat glands and, secreted into the sweat, and transported to the epidermal surface. This peptide fragment of DCD was named as DCD-1 which play a role in antimicrobial defence [277]. The prodomain, on the other hand, can be spliced to encode Y-P30 which results in 30 aa peptides which were previously named as diffusible survival evasion peptide (DSEP) and was shown to increase resistance to oxidative stress [277].

DCD function and localisation seems to mirror AGR2 since DCD is secreted and also found in extracellular vesicles, implicated in cancer cell survival. Therefore, DCD is a highly-relevant ‘dominant’ pro-oncogenic signalling target which can be a potential AGR2 client protein. Studying the interaction can also help us understand the role of AGR2 in secretion and trafficking pathways.

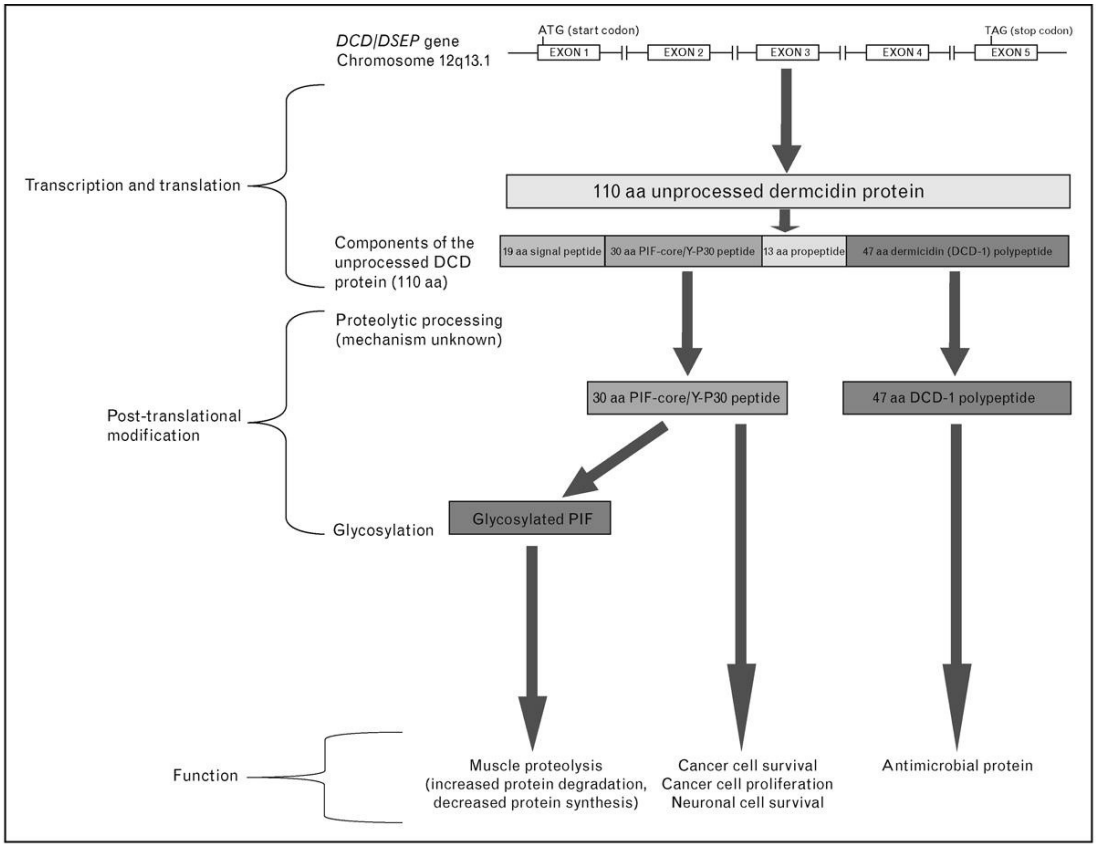


Figure 5.15 Overview of dermcidin gene and the gene products with distinct biological functions. (reproduced from [277]). Mature dermcidin RNA codes for a precursor protein with 110 amino acid residues and molecular weight of 11.2 kDa. DCD can be divided into two functional domains; Prodomain and antimicrobial peptide (AMP) domain. Prodomain can be spliced for example producing 30 aa peptides Y-P30 or 47 aa DCD-1 or 48 aa DCD-1L peptides. The resulting peptides have different biological roles in a cell.

5.2.6.1. Western blot analysis of AGR2-expressing cells to check DCD expression levels

We first assessed whether the DCD expression is indeed changed as a result of ectopically expressing AGR2. We repeated the transfection of RSV empty vector, wt-AGR2 and AGR2-KDEL into FLO-1 cells and incubated for 72 hours the same way we did for mass spectrometry analysis. The cells were harvested using urea lysis and subjected to western blot analysis. There were two peptides detected by mass spectrometry analysis and these are; dAVEDLESVGk and aVGGLGk where the former is a unique peptide. Both peptides attributed to AMP domain of the C-terminal DCD. Thus, we obtained the DCD antibody which recognises the C-terminal epitope of DCD (anti-Dermcidin E-7, #sc-393728, Santa Cruz Biotechnology). However, this antibody did not detect any band at the expected size of around 12 kDa even after long exposure and high concentration of lysates (data not shown). We then obtained a second antibody from a different commercial source which can detect endogenous DCD (anti-Dermcidin, #ab175519, Abcam) and also binds to C-terminal of DCD (aa 81-110). Immunoblotting using this antibody on the transfected cell lysates detected bands around 12 kDa and also a second upper band around 25 kDa which we assumed is the dimer form of DCD, and this band also appeared in manufacturer's datasheet. However, the corresponding 'monomeric' 12kDa band looks relatively higher in the control RSV empty vector expressing FLO-1 cells compare to wt-AGR2 or AGR2-KDEL cells. The 25kDA 'dimeric bands' appearance was more apparent, where in both wt-AGR and AGR2-KDEL cells, the bands were more reduced compared to the RSV control. This data contradicts with our mass spectrometry data suggesting that DCD is upregulated in both cases. We hypothesised that the DCD is secreted into media as DCD-1L/DCD-1 processed peptide form (the 48/47 aa long peptide respectively) during post-translational modification hence the lower expression was seen in western blot. Thus, we collected the media of the transfected cells, concentrated it 40 times and subjected to immunoblot analysis using DCD antibody. However, we could not detect any expression of DCD in the media (Figure 5.16B). Another explanation would be that, although fold changes in peptide ions identified from TMT quantitation can be used to translate protein changes as being "up or downregulated", this does not necessarily mirror changes in the absolute steady-state protein levels. The

fold changes might also reflect the extractability, post-translational modification, and trypsinisation capacity that give rise to peptide ion identification and then quantitation using MS. Since DCD is highly susceptible to post-translational modification (proteolytic processing to form functional peptides with as illustrated in Figure 5.15, this can affect quantitation using MS yielding such a result. Despite contradictory data (mass spec vs. western blot), the DCD level was altered as a result of AGR2 expression, suggesting that mass spectrometry was highly sensitive to detect any proteomic changes

5.2.6.2. AGR2 and DCD protein interaction using in situ proximity ligation assay

As mentioned previously, one member of the lab showed that DCD was one of the pulled-down proteins associated with ELMO1 protein as detected by SWATH MS (Ali Alsaadi Thesis 2016). Also, ELMO1 was shown to form protein-protein interaction with DCD and AGR2 as analysed using PLA. Thus, we evaluated whether AGR2 can form a protein complex with DCD in cells using PLA. We utilised the Duolink In-situ based PLA assay that we have used before to detect the AGR2 and EpCAM protein-protein interaction (Chapter 4).. We used the mouse monoclonal AGR2 antibody (AGR3 4.1) and rabbit polyclonal DCD antibody (Abcam). FLO-1 cells were first transfected with RSV empty vector (EV Upper panel), wt-AGR2 (middle panel) or AGR2-KDEL (lower panel) and incubated for 24 hours and processed for PLA using the antibody pair. Following addition of PLA probes, ligation and amplification, protein-protein interaction complex was detected with red fluorescent probes (Duolink). However, we could not detect any PLA foci in cells expressing either wt-AGR2 or AGR2-KDEL, suggesting that the two proteins did not interact with each other in cells (Figure 5.17, middle and upper panel). Based on these data, we hypothesised that AGR2 does not bind DCD directly but may bind through ELMO1 protein or its cofactors.

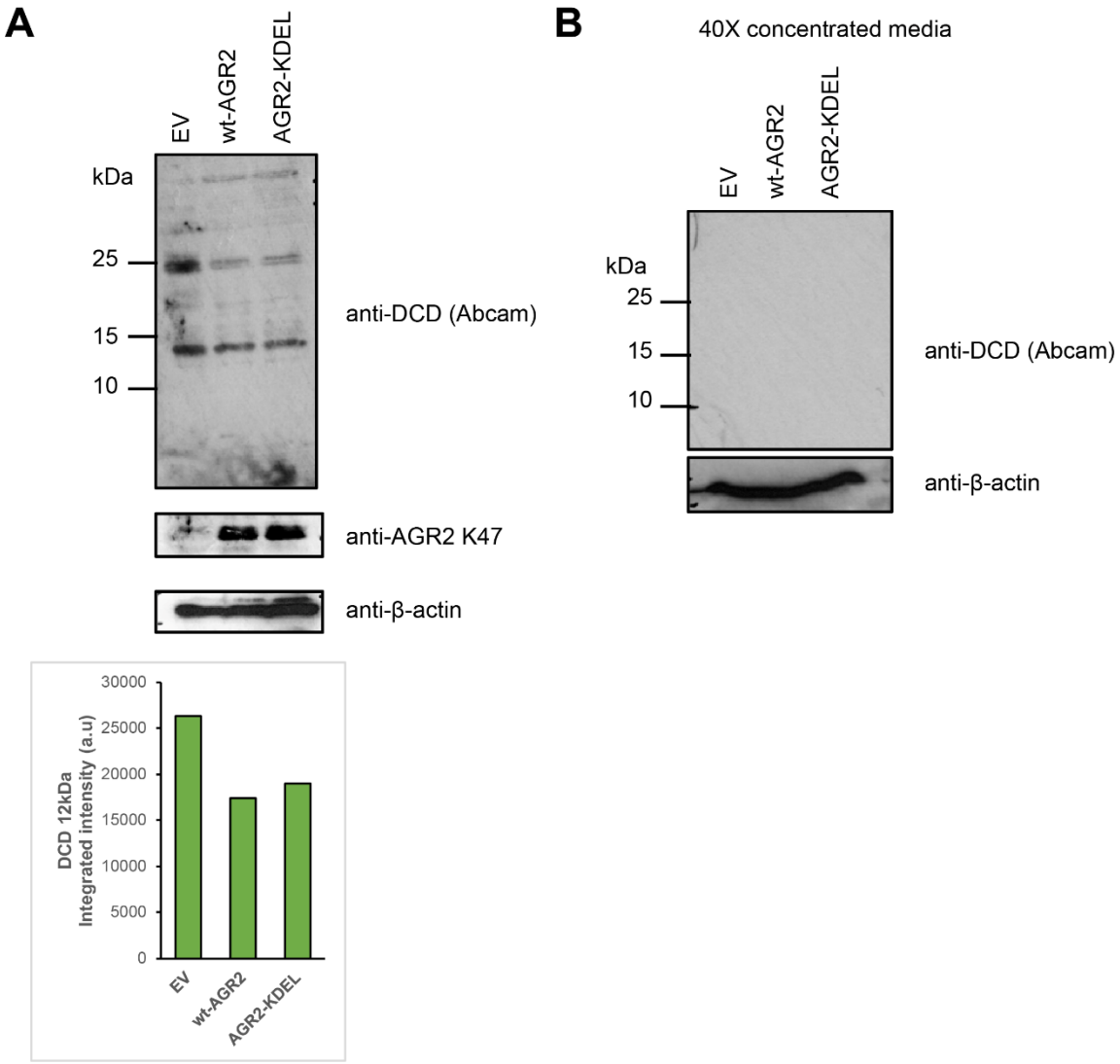


Figure 5.16 Western blot analysis of AGR2 reconstituted FLO-1 cells using DCD antibody. (A) (Upper panel FLO-1) Cells were transfected with RSV empty vector (EV), wt-AGR2 or AGR2-KDEL using electroporation and incubated for 72 hours, the same condition as our previous mass spectrometric analysis. The lysates were harvested using urea lysis buffer and immunoblot using anti-DCD (Abcam). (Lower panel) The 12kDa band intensities in all conditions were measured using ImageJ and plotted as a bar graph. (B) Media from cells transfected with constructs in (A) were collected and concentrated 40X using Amicon concentrating column with 30,000 da cutoff filter (Millipore). Concentrated media were mixed with SDS sample buffer and subjected to western blot using anti-DCD.

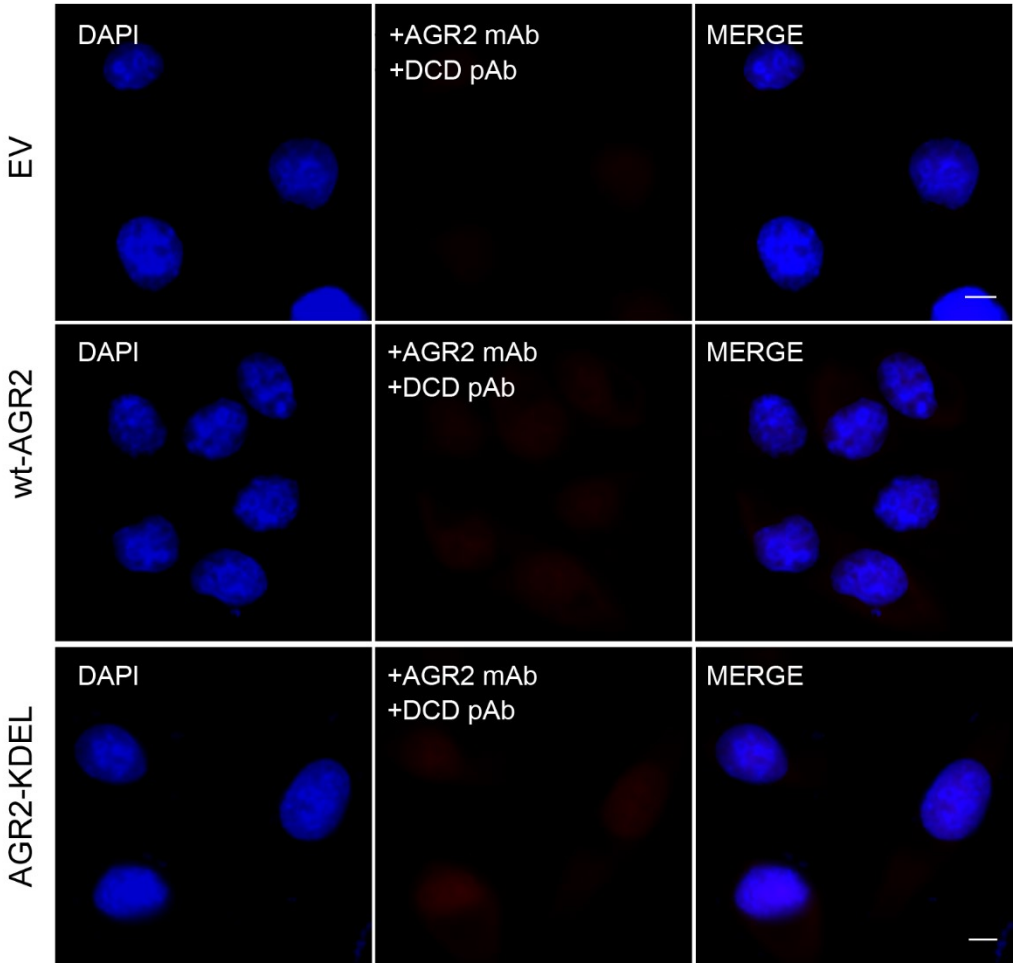


Figure 5.17 Proximity ligation assays demonstrate that AGR2 and DCD did not form protein-protein interaction *in situ*. Representative image of a proximity ligation assay performed with antibody pair of anti-AGR2 mAb (AGR3 4.1) and anti-DCD pAb (Abcam). FLO-1 cells were first transfected with RSV empty vector (EV Upper panel), wt-AGR2 (middle panel) or AGR2-KDEL (lower panel) and incubated for 24 hours before processed for PLA using the antibody pair. PLA probes (Anti-rabbit PLUS probe and anti-mouse MINUS probe) were then added to the samples. Following ligation and amplification, protein-protein interaction complex was detected with red fluorescent probes (Duolink). Red fluorescence foci indicate the interaction between the two proteins, which was not observed in any of the cells. Scale bar 10 μm.

5.3. Discussion

Methods to study protein-protein interactions have been improved over the years. A typical protein-protein interaction study usually begins with an initial screen for novel binding partners. It has been shown previously that knockout of a protein which has a central role in many networks tends to be lethal [281]. This phenomenon has been observed in many organisms [282, 283] and is commonly referred to as the centrality-lethality rule. The current view that most, if not all, cellular proteins are directly or indirectly coupled to a large cellular protein–protein interaction network has implications for the way we define cellular pathways.

The emerging programmable nucleases for a precise genome editing has enabled scientists to produce rapid and complete loss-of-function gene knockout in cells [268, 270, 284] to study gene function. In this chapter, we aimed to develop an isogenic cancer cell line by knocking out AGR2 using the latest CRISPR/Cas9 genome editing technology to infer AGR2 function and use the AGR2-null cell to screen for global proteome changes using quantitative proteomics. We have previously shown that the database mining of proteins containing AGR2 docking motif shows enrichment in transmembrane proteins (Chapter 3), suggesting that AGR2 is involved in the maturation of this class of protein. Further analysis of AGR2 with one of the proteins of this class, EpCAM showed that EpCAM's plasma membrane expression is dependent on AGR2 expression. Immunohistochemistry using OAC tissues showed that the two proteins were highly co-expressed (Chapter 4). We aimed to study AGR2 function by a complete loss-of-function which is then coupled with quantitative proteomics analysis to identify the dominant pathways or co-factors associated with AGR2. Our goal was also aimed to find evidence that AGR2 engages in trafficking and secretion pathways to complement our previous data suggesting that AGR2 involve largely in the maturation of proteins destined for secretory pathway.

We first attempted to generate a complete loss-of-function by knocking out AGR2 in MCF-7 cell line using CRISPR/Cas9 technology. We targeted the exon 4 of the AGR2 genomic locus for DSB. However, we were unable to generate AGR2

knockout cell line as confirmed by western blot analysis and DNA sequencing (Figure 5.4and Figure 5.6). A different experimental strategy was employed by using a different algorithm for designing the 20-nt gRNA, a different cell line, and FACS sorting to isolate individual clones yield the same results, where no clonal line carrying AGR2 knockout was isolated (Figure 5.7). However, in both experimental approaches, we observed variable AGR2 expression in the clonal lines compared to the parental cell line, suggesting that there were selection pressures during single cell dilution and expansion that can give variable gene/protein expression levels (Figure 5.4). There was no published data thus far showing AGR2 knockout in a cell line by the CRISPR/Cas9 technology though partial knockdowns by RNAi have been widely successful [144, 161]. There are several aspects that affect CRISPR/Cas9 technology and this includes efficiency and specificity, including Cas9 activity, target site selection and sgRNA design, delivery methods, off-target effects and the incidence of NHEJ/HDR repair. One of the main problems we encountered was the delivery of the plasmid encoding the sgRNA and Cas9 since different cells have different transfection efficiency. We utilised both transfection and electroporation methods coupled with FACS sorting to isolate clonal lines. The electroporation method for example only yield 20% efficiency as counted by FACS for cells with GFP staining. Additionally, once transfected, plasmid DNA can also persist inside the cells for several days, which may exacerbate off-target effects in this small population of transfected cells. To circumvent these pitfalls, studies have emerged to improve this technology for example, by improving sgRNA design by extending PAM sequence [285]. Moreover, the delivery method by viral vectors have been used as this vector is an active, receptor-dependent process that allows control over the amount of DNA transfected with greater efficiency and less cytotoxicity and viral vector can enter into large numbers of cell types efficiently [286]. Researchers are also looking to microbes to find endonucleases that have different sequence requirements than Cas9 so that they can expand the flexibility in modifying the number of sequences. One study has found an endonuclease which is called Cpf1 [287]. This enzyme is smaller than Cas9, has different sequence requirements and is highly specific. Another enzyme, called C2c2 has been found that can targets RNA rather than DNA and holds the potential to develop into new RNA-targeting tools [288].

The inability to knockout AGR2 suggests that AGR2 may play a central role in the cell, and complete loss-of-function of the gene may be lethal. This observation was seen in mice model, where conditional AGR2 knockout (*Agr2 -/-*) resulted in premature death due to gastrointestinal obstruction showing consistent phenotype featuring hyperplasia of the glandular stomach [289]. Two other studies using conditional (*Agr2 -/-*) null mice experiments showed that the mice were viable but with histologic changes, if not severe changes, in the stomach and intestines [60, 63]. The different results from *Agr2* null mice may be due to conditional knockout that implements floxed *Agr2* exons whose deletion were dependent on the constitutive expression of Cre recombinase. In addition, there might be some residual AGR2 expression at early developmental stages, hence the increased in viability. A more recent study by Dong *et al*., generated a non-conditional AGR2 null mouse with germline deletions of *Agr2* exons 2–8 [111]. The study showed that only two out of 52 mice with homozygous AGR2 knockout is viable but one of them died after 3 weeks while the other suffered from weight loss and displayed many of the gastrointestinal obstructions. This suggests that AGR2 may belong to highly connected proteins in the cell that is important for its survival, and also explains why we were unable to generate AGR2 knockout in cell line using CRISPR/Cas9 technology. AGR2 was found to be overexpressed in our proteomics screen in tissues of OAC compared to normal adjacent tissues (Robert O'Neill Thesis 2014) and confirmed using immunohistochemistry of tumour microarray earlier in the thesis (Chapter 4) signifying that AGR2 is needed for cancer cell survival. Inducible knockdown of AGR2 in a cell line has never been done before. Inducible knockdown RNAi System may be performed in the future to determine the lethality of AGR2 knockdown in cells as this has been proven useful in cases where suppression of a gene may be lethal, preventing its analysis [290] and this can support our current hypothesis. This system use an expression vector for example, lentiviral expression system to express the and this system is induced when either tetracycline (Tc) or doxycycline (Dox; a Tc derivative) is added to the culture medium of an appropriately engineered cell. Induction of the shRNA results in suppression of the gene targeted by the shRNA through RNAi. Thus, the system allows for tight regulation of the expression of a

shRNA in a time-dependent manner suitable for studying lethality of a gene when Tc or Dox is added to the medium.

Our attempt to generate an overexpressed AGR2 by creating a stable cell line expressing AGR2-GFP were also unsuccessful. Real-time visualisation of transfected AGR2-GFP in both cell lines showed that AGR2-GFP expression declined to almost non-existent after about 24 hours post-transfection (Figure 5.8). As AGR2 protein has been shown to promote cell adhesion [61], we hypothesised that AGR2 expression stimulates 'adhesive properties' which can prevent cells to proliferate. However, when AGR2-CHERRY was subjected to live-cell fluorescence imaging in FLO-1 cells, the expression was observed as puncta or secretory vesicles which implied that the AGR2-CHERRY can be expressed in this AGR2-null cell. These vesicles move rapidly throughout the cells suggesting that AGR2 not only reside in the ER, but can escape or 'bud off' from the ER to be secreted or perhaps involve in trafficking of secretory proteins as we hypothesised (Figure 5.9).

This led us to find a new strategy to study AGR2 function in which we utilised protein overexpression strategy to study AGR2 function. We utilised AGR2-negative cells, FLO-1 that we identified earlier in which we reconstitute the cell as a function of AGR2 transient overexpression and process for mass-spectrometry analysis. Overexpression of AGR2 can promote changes in the AGR2-interactome by either dissociating existing PPIs or induction of new AGR2 interactions especially in cells that do not express AGR2. We delivered the weaker expression of AGR2 under the control of Rous sarcoma virus (RSV) promoter into FLO-1 cells, which was one of the commercially weakest promoter previously cloned (Chapter 4) to ensure that its expression is not too overwhelm that can be toxic to cell and also warrants the true representative of AGR2-reprogrammed cells for our proteomics screen analyses. Quantitative proteomics mass spectrometry was used to look for a global view of proteomic dynamics to unravel dominant proteins and pathways of wt-AGR2 and AGR2-KDEL expressing FLO-1 cells compared to RSV empty vector FLO-1 cells. Progress in mass spectrometry instruments and data analysis currently can be carried out using both label and label-free approaches to quantify the protein expression

landscape of the experimental condition (Table 5.2 Summary of quantitative proteomics strategies). Quantitative proteomics can be split into two forms: relative and absolute. Relative quantitation methodologies compare two or more samples to one another (e.g to a control) as a function of relative fold change of protein abundance. Absolute quantitation is the determination of the precise amount or mass concentration of a protein. In its simplest form, 2-dimensional gel electrophoresis (2-DIGE) coupled to fluorescent dyes, fluorophores or radioactivity label allow protein levels to be quantified using the intensity of the incorporated labels to the same spot from a control lysate. In MS-based proteomics, the reliance is based on stable isotope dilution theory, which states that a stable isotope-labelled peptide is chemically identical to its native counterpart; therefore the two peptides also behave identically during chromatographic and mass spectrometric analysis. The sensitivity of an MS detector means that the mass difference between the labelled and unlabelled isoforms of the peptide can be differentiated, and quantification can be achieved by comparing their relative signal intensities. Isotope labels can be introduced into the sample in several ways: (i) metabolically, (ii) through chemical reactivity, (iii) enzymatically or (iv) as an external standard of the spiked synthetic peptide to normalise to [291]. Recently, label free strategies have begun to emerge, such as spectral counting where the number of spectra acquired of a specific peptide/protein is used as an estimate of abundance in the sample. The stable isotope labelling by amino acids (SILAC) method for example use the metabolic labelling of cells in growth medium [292] incorporating light and heavy forms of arginine or lysine allowing labelling of all nascent proteins in the experimental condition. This makes proteins from the light and heavy cells distinguishable by MS after the cell lysates are mixed and the proteins separated. However, multiplexing is limited in SILAC in which three isotopically labelled SILAC allows comparison of three cell populations in a single experiment [293]. SILAC also is only restricted to cultured cells and depends on full incorporation of amino acids in growth medium for mass spectrometric analysis. The development of isobaric mass tags such as tandem mass tag (TMT) [272] and isobaric tags for relative and absolute quantification (iTRAQ) [273] allows for the comparison of up to ten samples in parallel, which is particularly useful in studies of time courses or in providing biological replicates. TMT used in this study is based on the introduction of isobaric

tags to label peptides which do not influence chromatographic nature. The LC-MS/MS analysis yields peptide fragment ion spectra generated in MS1 for peptide identification and the cleaved tag spectra generated in MS2, which are used for relative quantitation. The ratio of detection of the tags can then be used as a relative quantitation measure to detect the difference between the levels of proteins from different samples. Assuming complete peptide labelling with TMT isobaric tags of each sample, the more abundant peptides within each sample will incorporate more label. When equal amounts of each sample are mixed and subjected to LC-MS/MS analyses, those samples with a higher concentration of peptide will produce higher reporter ion peak intensities in the MS2 scan. By comparing the relative reporter ion intensities, the relative peptide and therefore protein abundances in the original samples can be determined [294]. Therefore TMT methodology provided the most appropriate platform for this study as we can perform the TMTsixplex™ experiment which permits comparison of not only AGR2-null and AGR2-expressing cells in biological duplicates, but also the comparison with the enhanced ER-localised AGR2, AGR2-KDEL to control for ER/intracellular specificity.

TMT relative quantitation showed differences in the proteome dynamics of wt-AGR2 and AGR2-KDEL expressing cells compared to RSV empty vector cells (Figure 5.12). The overall protein expression levels quantified across the two conditions were not significantly altered as the fold changes were largely between 0.5-1.5 region (Figure 5.12A-B). However, the proteome of reconstituted wt-AGR2 cells showed that most proteins were downregulated while the AGR2-KDEL cells showed that most proteins were upregulated suggesting that the two AGR2 variants have specific distinct signalling pathways. This also suggests that the localisation of AGR2 gives different cellular reprogramming outcomes. The wt-AGR2 has been shown to localise intracellularly (mainly in the ER) as well as extracellularly [94, 260]. The AGR2-KDEL, with more classic ER retention motif, was shown to localised intracellularly, more specifically in the ER where it colocalises with ER-marker PDI [260]. A large number of downregulated proteins identified in the wt-AGR2 expressing cells may correlate to our real-time analysis (Figure 5.8), where the wt-AGR2-GFP expression in FLO-1 cells was not well- maintained or was almost non-

existent in FLO-1 cells. Possibly, the wt-AGR2 is secreted to an extracellular milieu which can play a role in autocrine/paracrine signalling giving such results. The enhanced ER-localised AGR2-KDEL, on the other hand, may act as general molecular chaperone, assisting in the proper endogenous protein folding and maturation in the ER, yielding many proteins exiting the stringent ER quality control, hence many upregulated proteins quantitated by mass spectrometry.

In Chapter 3, AGR2 has propensities to bind transmembrane proteins as this type of proteins were found to be dominant using our motif-based approach. This leads to another hypothesis in this chapter where proteomics approach could show that membrane proteins or perhaps secretory proteins may increase globally when AGR2 is expressed. However, our results did not show dominant upregulation of membrane proteins (Figure 5.13). We hypothesised that the transmembrane proteins were not properly extracted during sample preparation due to its hydrophobic and water-insoluble nature, therefore mainly soluble proteins were extracted for mass spectrometric analysis. This experiment can be repeated in the future by performing an enrichment and fractionation of subcellular compartments prior to mass-spectrometric analysis to increase the detection limits of analysis and this also can determine the localisation of identified proteins. In our case, the membrane fraction can be extracted for mass spectrometric analysis to check whether membrane proteins are upregulated in the presence of AGR2 to support our hypothesis. Alternatively, proteome analysis of a specific subcellular fraction that has been post-translationally modified for example, protein glycosylation-targeting enrichment can be performed. Protein glycosylation analysis has become an important target in the proteomic research field and has great potential for clinical applications like the development of biomarkers. Many protein biomarkers that are known to be cancer biomarkers are also glycoproteins [295, 296] and importantly we found EpCAM which is a type of glycoprotein.

Methodology	Labelling strategy	Quantitation	Advantages	Disadvantages
2-DIGE	Spot analysis or fluorescent/dye labelling	Relative	Simple, measures whole protein	Limited range of coverage
iTRAQ TMT	Isobaric labelling	Relative	Allows multiplexing up to 10 samples	Additional chemical modification step
SILAC	Heavy and light metabolic labelling	Relative	Concurrent processing of samples	Requires incorporation in cultured cells and limited to in in-vivo use
^{18}O Labelling	Heavy oxygen isotope labelling	Relative	Labelling during trypsin digestion	Complete labelling is rare and poor coverage
Spectrum count	None	Relative	No labelling	Inaccurate for low abundance
AQUA	Spiked heavy peptide from known protein	Absolute	Absolute quantitation	Only suitable for known targets

Table 5.2 Summary of quantitative proteomics strategies

Functional analysis of the top 40 upregulated proteins in both AGR2 variants expressing cells showed that they were enriched in exosomal proteins (Figure 5.14). This result is consistent with previous studies that showed AGR2 upregulates exosomal marker TSG101 whose expression induced p53-dependent activity suppression whose expression attenuates p53 signalling [297], and strongly suggest that AGR2 can sculpt cellular secretome. The top overexpressed proteins are also enriched in lysosomal proteins. Lysosomes are membrane-bound organelles that receive and degrade macromolecules from the secretory, endocytic, autophagic and phagocytic membrane-trafficking pathways through the activity of lysosomal hydrolases [298]. Additionally, lysosome also has been linked with various physiological processes, such as cholesterol homoeostasis, plasma membrane repair, bone and tissue remodelling, pathogen defence, cell death and cell signalling, suggesting possible roles of AGR2 in any of these processes. The AGR2-KDEL, but not wt-AGR, expressing cells were enriched in cytoplasmic proteins, which is consistent with our hypothesis that ER-localised form of the AGR2 act as a protein folding catalyst releasing properly folded protein from ER to cytoplasm. There are 15 common proteins in the top 40 upregulated proteins suggesting possible AGR2 client proteins. One interesting protein include superoxide dismutase 1 (SOD1) which is upregulated in both wt-AGR2 and AGR2-KDEL samples, also referred to as Cu/Zn

superoxide dismutase, which is an antioxidant enzyme that binds copper and zinc ions and forms a homodimer whose main known function is as a dismutase removing toxic superoxide radicals (O_2^-) radical into either molecular oxygen (O_2) or hydrogen peroxide (H_2O_2). SOD1 mutations have been associated with the neurodegenerative disease amyotrophic lateral sclerosis (ALS) [299]. Since AGR2 has an oxidoreductase activity through its thioredoxin motif, it is interesting to explore how SOD1 and AGR2 control the cellular redox state, particularly in the ER. The increased expression of SOD1 in both wt-AGR2 and AGR2-KDEL suggested that the levels of oxidised proteins are elevated. However, this theory is not tested and the interaction of AGR2 and SOD1 can be validated in the future.

We chose DCD protein as potential AGR2 interacting protein as this protein was upregulated in both AGR2 conditions. The fact that DCD is secreted and implicated in cancer cell survival serves the protein as a highly relevant AGR2 client protein candidate. Studying the interaction can also help to identify cofactors involved in secretion and trafficking pathways. There were two peptides detected in our mass spectrometry analysis, and these peptides belong to the C-terminal AMP domain, and these peptides have homology with all 13 known processed AMP peptides (Figure 5.18). AMP domain is responsible for a broad range of antimicrobial activity against pathogens and was constitutively secreted in eccrine sweat. We analysed the cells transfected with wt-AGR2 and AGR2-KDEL using western blot, to see if the DCD is indeed elevated in these reconstituted cells. However, we have difficulties in finding a good antibody to detect the presence of DCD. We obtained two antibodies that detect the C-terminal of DCD. The first antibody from Santa Cruz Biotechnology did not detect any band in western blot. The second antibody detects the expected full-length DCD at 12 kDa and also a ‘dimeric band’ (Figure 5.16). The expression of DCD as detected by the relative intensities of both the 12kDa and dimeric band showed that the empty vector control has elevated expression compared to the AGR2 expressing cells. This is inconsistent with our mass spectrometry data which suggest that DCD is more abundant in both AGR2 variants expressing cells (Figure 5.12C). Western blot analysis of the concentrated spent culture media from AGR2 transfected cells did not detect the presence of secreted DCD (Figure 5.16B). Despite the contradictory data,

mass spectrometry quantitation showed that the DCD is perturbed in both wt-AGR2 and AGR2-KDEL expressing cells proving that mass spectrometry is sensitive to detect proteomic changes. The contradictory data (MS vs western blot) has been observed in our lab recently in a proteomic screen study to detect changes identify Nutlin-3 responsive MDM2-binding proteins [300]. Thus, the fold changes may not necessarily reflect protein abundance but may reflect post-translational modification (in our case). DCD has a high susceptibility for proteolytic processing forming functional peptides with distinct biological roles. In eccrine sweat, 14 proteolytically processed DCD peptides are found, one peptide derived from the N-terminal prodomain named YDP-42 (can be further processed to PIF or Y-P30) and 13 peptides derived from the C-terminal AMP domain (Figure 5.18). The AMP domain is derived from the C-terminal region of the precursor protein, DCD-1L (consisting of 48 amino acids) and DCD-1 (47 amino acids, lacking the last leucine) and this can be further processed by post-secretory processing. Thus, selection of the antibody must be highly specific depending on the peptide of interest, and the lack of antibodies detecting all peptides makes detection difficult. It is also possible to check mRNA level of DCD in the future as upregulation of DCD at transcriptional level may suggest upregulation of DCD at protein translational level. This could be tested by using quantitative real-time PCR (qRT-PCR) comparing DCD mRNA expression levels in cells with AGR2-transfected cells and AGR2-negative control cells. If the mRNA is indeed upregulated, this suggests that DCD protein level is subjected to the post-translational modification that is difficult to detect.

DCD was shown not to interact with AGR2 in cells as tested using PLA (Figure 5.17). DCD was also found in another proteomics screen in the lab using SWATH mass spectrometry analysis of ELMO1 pull-down assay from FLO-1 and OE19 cells (Ali Alsaadi thesis 2016, and personal communication). ELMO1 was shown to form protein-protein interaction complex with AGR2 as well as DCD in cells using PLA. Since we do not have evidence that AGR2 interacts with DCD, we speculate that AGR2 may bind DCD through the interaction with ELMO1 or with one of ELMO1 signalling proteins. It remains to be determined the effect of AGR2-induced DCD expression and secretion in tumour development. It is also interesting to establish

whether AGR2 contributes to the sensitivity of proteolytic processing of DCD. The crystal structure DCD-1L peptide from the AMP domain has been established showing that the peptide exists as hexameric bundle formed by elongated α-helices and stabilised by the presence of zinc [301]. DCD forms highly efficient channels to disrupt the bacterial transmembrane potential that is essential for cell survival. Therefore, it is fascinating to explore in the future if there is any relationship of AGR2 in the oligomerization of DCD-1L peptide to form the ion channels and its contribution in the antimicrobial barrier.

Figure 5.18 Processed peptides identified from the different domains of the dermcidin precursor protein. (Adapted from [280]) The dermcidin precursor protein has a high susceptibility for proteolytic processing. After cleavage of the 19-amino-acid *N*-terminal signal peptide and secretion into sweats, the 91-amino-acid dermcidin precursor protein can undergo further proteolytic processing, generating 13 functional peptides.

CHAPTER 6: Developing Synthetic Models to Dissect AGR2 Function

6.1. Introduction

6.1.1. Antibodies and the emerging of recombinant antibodies

Antibodies are specialised immune glycoprotein secreted by B cells which are produced by the mammalian immune systems to protect against foreign antigens such as pathogenic organisms or toxins and has the ability to neutralise them. Also known as immunoglobulins (Igs), can be divided into two categories: polyclonal (pAbs) and monoclonal antibodies (mAbs). Polyclonal antibodies are a collection of antibodies from different B cells recognising multiple epitopes on the same antigen, while monoclonal antibodies are derived from a single B cell lineage and therefore only bind to one unique epitope of the target antigen. Thus, mAbs promise high-specificity compared to pAbs counterpart in which pAbs may possess cross-reactivity to different biomolecules with similar epitopes. Consequently, mAbs have been the highly-favoured pharmaceutical molecules and are the fastest growing component used in immunotherapy identifying targets in wide-range of diseases and conditions [302] .

Antibodies come in different forms known as isotypes. There are five antibody isotypes in mammals: IgM, IgD, IgG, IgA, and IgE [303]. They differ in the type of heavy chains that the antibody possesses, with each heavy chain class named alphabetically: α (alpha), γ (gamma), δ (delta), ε (epsilon), and μ (mu). IgG is the most abundant form of immunoglobulins, and it provides the majority of antibody-based immunity against invading pathogens. IgG also is most extensively used molecule in therapeutic antibodies [304]. Moreover, it is the only class that crosses the placenta and is further subdivided into 4 different classes: IgG1, IgG2, IgG3 and IgG4 which confer distinctive functional properties at the C-terminal part of the heavy chain. All antibody molecules, however, are constructed the same way, having a Y-shaped signature consisting of four polypeptide chains; two identical heavy chains (V_H) and two identical light chains (V_L) connected by disulphide bonds (Figure 6.1, top panel). Each chain is further classified into variable and constant domains. Antibodies are large molecules with a size of approximately 150 kDa. The variable region (Fv) amino

acid sequence varies greatly among different antibodies, and it is located in the Y tips. This variable region, composed of 110-130 amino acids, give the antibody its specificity by serving as an antigen contact site making each antibody molecule at least bivalent. The variable region is further subdivided into complementarity determining regions (CDRs) and framework regions (FR). The constant region (Fc) determines the mechanism used to destroy antigens such as binding to various cell receptors, activation of the classical complement pathway, and phagocytosis [305]. Soon, it was discovered that the structure is not suitable for some applications due to some effects triggered by the Fc domain. For instance, Fc-mediated effects can lead to toxicity due to cytokine release and have a long serum half-life results in reduced discrepancy in imaging applications [304]. Initially, researchers attempted to remove Fc domain by using proteolytic treatments with enzymes such as papain or pepsin. This yield fragment antigen binding (Fab) or two Fab arms linked together F(ab')2 fragments (Figure 6.1, bottom panel) [306, 307]. The cleaved Fc domain, however, does not have any variation and triggers the biological effector functions of the antibody [308]. Later, advances in recombinant DNA technology and genetic engineering directed the development of a large variety of recombinant antibody (rAb) fragments with unlimited potential for research, diagnostics and therapy, and have emerged over the last two decades as the fastest growing class of therapeutic proteins [302]. The fragments include single chain variable fragment (scFv), V-domain molecules as well as camelid V_HH and shark V-NAR fragments (Figure 6.1, bottom panel). The minimised monoclonal antibodies retain specificity as the full-length IgG. In addition, they offer several advantages such as better tissue penetration, rapid blood clearance and tumour uptake [304]. The rAb fragments can also be easily and cost-effectively cloned and expressed in large quantities in a wide range of hosts [309]. This has led to the new generation and innovative antibodies that have been implemented in research studies and therapeutics such as antibodies coupled to radioactive tags for cell imaging, therapeutic ligands, enzymes and pro-drugs for treatments such as Antibody Dependent Enzyme Prodrug Therapy (ADEPT) [310]. The mAb-fragment repertoires can be applied and manipulated to better suit the future application of the fragments such as the proteomic discovery of new cancer biomarkers, and development of nanosensors for the diagnostic purpose [304].

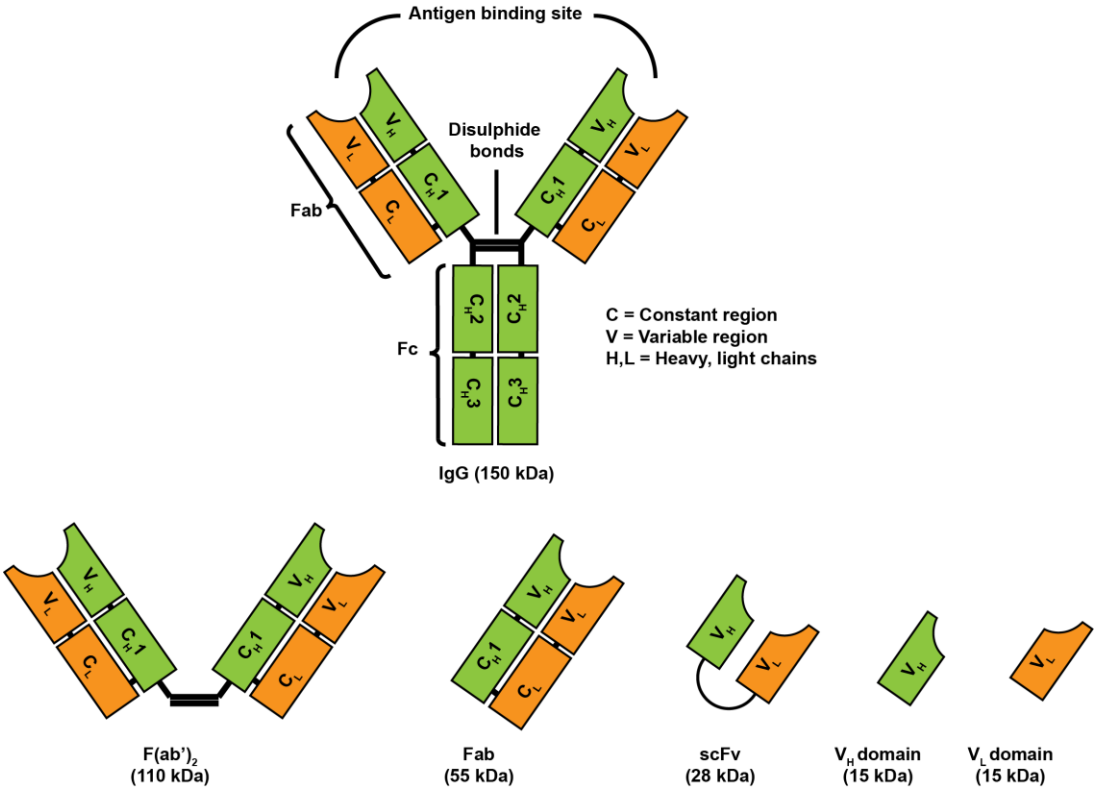


Figure 6.1 Schematic representation of antibody structure and antibody fragments. (Top panel) The basic structure of antibody molecule represented by full-length IgG consists of the Y-shaped structure. The molecule is composed of two identical heavy and light chains. Each of these chains contains multiple constant (C) and one variable (V) regions linked by disulphide bonds. The antigen binding site is located in the N-terminal of variable domains of both chains. Variable domains contain complementary determining regions (CDRs) and framework regions (FR) that highly assist in antigen recognition. Heavy chain constant domain is divided into three parts i.e. CH1, CH2 and CH3 where both CH2 and CH3 from the Fc portion of the antibody responsible for its effector functions. (Bottom panel) Antibody fragments resulted from proteolytic digestion producing Fab and $F(ab)_2$ region containing variable and constant domains from heavy and light chains with no Fc portion of an antibody. Also shown scFv containing only variable heavy and light chain domains connected by a flexible linker, and single-domain containing heavy or light chain.

6.1.2. Phage display

Antibody generation has been dramatically accelerated by *in-vitro* selection systems, particularly phage display. Phage display technology is a commonly used method for the selection of monoclonal antibodies against various antigens. The concept of fusing foreign DNA fragments with gene encoded for pIII coat protein of a non-lytic filamentous phage was first revealed by Smith in 1985 [311]. Five years later, McCafferty *et al.* established for the first time that scFv could be displayed on the surface of the phage surface as a functional protein retaining its antigen binding capacity [312]. Since then, the technology has been applied for the selection of diverse molecules such as Fab fragments, polypeptides, peptides, zinc fingers as well as enzyme modules. The display of the monoclonal antibodies on a filamentous bacteriophage allow rare clones to be screened and isolated from a large-scale phage library against any antigen of interest. A single phage library can contain up to billions of distinct clones. Phage and phagemid are two vectors used for expression, where the latter is most commonly used. Phagemid requires additional proteins called helper phage for packaging necessary gene products to make functional phage particles displaying recombinant protein on the surface of a phage.

There are three types of phage display libraries which are categorised based on the origin of antibody genes; immune, naïve, and synthetic libraries [309, 313]. Immune libraries are constructed from variable domain originated from B cells obtained from immunised animals such as mice, sheep, rabbit, camel as well as humans with the target of interest. As compared to a library made from hybridoma cells, library made from spleen might give better diversity but also contain non-specific antibodies. This library can result in high-affinity selection of antibodies as it is biased only towards the antigen of interest used for immunisation. Thus, a new phage library has to be constructed for each antigen which can be very time-consuming. Secondly, naïve libraries are created from B cells of non-immunized donors. Unlike an immune library, this library has the advantage that it can be used for a large number of antigens and beneficial for the construction of antibody fragments which are difficult to generate using hybridoma technology. The affinity of an antibody towards a specific target is thought to be proportional to the size of the library [314]. Construction of variable

domains involve PCR amplification from IgM mRNA or total mRNA, as building libraries from IgG mRNA do not perform well due to unrelated immune response [315]. Thirdly, synthetic libraries are also derived from the nonimmune donor but are built artificially by the *in-vitro* assembly of heavy and light chains gene segments and introducing complementarity determining the region (CDRs) of varying loop lengths using PCR and randomised oligonucleotide primers [316]. Most structural and sequence diversity is found in the CDR3 region of the heavy chain and essentially responsible for antigen contact, with the other CDRs possessing limited variation [317].

6.1.3. Aims of chapter

Based on our prior report using biochemical approaches (Chapter 3) and the cell system analysis (Chapter 4), we hypothesised that AGR2 functions to elevate EpCAM protein levels occur via docking to the TxIYY binding site on the EpCAM receptor. The objective of this chapter is to deconstruct this hypothesis using synthetic protein modules. We wanted to explore if we could isolate synthetic scFv antibody fragments that bind to AGR2 with high-affinity from a massively diverse phage-display library, express the isolated scFv targeting AGR2 in mammalian cells and assess whether such high-affinity binding protein positively or negatively regulated by AGR2. We also aim to exploit the AGR2-binding peptide motif (TxIYY) previously isolated from peptide phage-display by making a synthetic AGR2-binding mini-protein, consisting the TxIYY docking site motif and allow its destination to reside in the endoplasmic reticulum and test the fate of such an AGR2-binding protein. We also seek to generate synthetic membrane protein, which contains a transmembrane domain and AGR2-binding motifs TxIYY. This provides an optimal 'artificial AGR2 client protein' and test the fate of such a membrane protein (e.g whether it can be regulated by AGR2 or whether AGR2 has any influence in the delivery of the synthetic membrane protein to the cell surface).

6.2. Results

6.2.1. Development of scFv antibody fragments targeting AGR2

6.2.1.1. scFv antibody fragments format for selection of monoclonal antibodies targeting AGR2

ScFv fragments represent the smallest functional antibody in which the V_H and V_L domains are joined with a flexible polypeptide linker preventing dissociation and retain the ability to bind to the antigen in high-affinity. It was first developed by Huston and co-workers [318] and refined by Whitlow group [319]. Nowadays, scFv is the most popular rAb fragment due to their versatility. These antibodies are much smaller in size ranging from 26-30 kDa and still retain affinity to bind target antigen [319]. The scFv antibody consists of one V_H and one V_L chain connected by a flexible peptide linker. The orientation could be in either V_L-linker-V_H or V_H-linker-V_L orientation, but the latter is the most common. The orientation can influence scFv stability binding to the antigen [320] and expression efficiency [321]. The length of the flexible linker also can affect the correct folding of scFv proteins and it was estimated that the linker is optimum at 3.5 nm resolution [322]. The most extensively used linker was the one that made up of glycine and serine stretch residues which offer great flexibility and combination with glutamic acid and lysine can increase solubility [323]. As scFv fragments bind monovalently to their target and they often present in low functional affinity (also termed avidity) and a short *in-vivo* half-life [324]. These properties can be useful for diagnostics imaging applications, for example, but they can compromise the efficacy as therapeutics agents as these may require higher retention times on the target antigen or engagement of multiple receptors in order to activate signal transduction and/or apoptosis [325, 326]. In order to overcome this drawback, scFv antibodies are engineered into different types of multimeric complexes for greater binding avidities and improved pharmacokinetic properties [327, 328]. The multimeric complexes include minibodies, diabodies, triabodies, tetrabodies and bispecific scFv fragments. ScFv antibody fragments can be selected and affinity matured by three different molecular displays, namely phage-display [312], ribosome display [329] and cell surface display [330, 331]. All three formats share the same basic principle in term of library construction of display libraries, which are created by cloning a massive collection of rAb genes and these libraries are screened for target antigen binding. Multiple selection

rounds of the polyclonal pools are performed to enrich for high-affinity monoclonal scFvs binding to target antigen. Here, we used phage display format to select for high-affinity scFv antibody fragments binding to AGR2 (Figure 6.2).

6.2.1.2. Isolation of high-affinity scFv antibody fragments binding to AGR2 from phage-display library

AGR2 was previously purified using affinity chromatography containing N-terminal His-tagged epitope (Chapter 3) and used as an antigen to select for high-affinity monoclonal scFv antibodies. The resulting His-AGR2 protein was highly pure showing a major monomeric band at 20kDa and a minor dimeric band as detected by coomasie blue staining and also immunoreactive in western blot analysis. His-AGR2 or coating buffer (negative control) was coated onto microtitre wells and a phage-scFv library was used to isolate individual AGR2 binding scFv fragments (Figure 6.2) in a screening process called biopanning. Briefly in biopanning, a phage display library is constructed which involves inserting foreign desired gene segments into a region of the bacteriophage genome, so that the peptide or antibody product will be displayed on the surface of the bacteriophage virion. The library is then incubated with a target molecule. The library can be incubated directly with an immobilised target or pre-incubated target prior to capture on a solid support. As in affinity chromatography, only specific peptides or antibodies are bound to the target while the non-interacting peptides or antibodies are washed away. The interacting phage display peptides or antibodies are then eluted and can be amplified by bacterial infection to increase their copy number. This screening and amplification process can be repeated several times to enrich higher-affinity phage display peptides or antibodies. The scFv library was constructed by cloning scFv genes into phagemid vectors by PCR amplification of RNA from a canine naïve library derived from canine spleen, and these were provided by Dr Euan Murray (Prof. Hupp lab). This library encodes both the heavy (V_H) and light(V_L) chains ligated together and connected via a linker of 15 amino acids, forming scFv fused to a gIII protein expressing on the surface of M13 bacteriophage (Figure 6.3). For the initial round of panning, the phage-scFv library was pre-cleared on a non-targeted His-tagged protein Reptin (a kind gift from Lucy Remnant) to exclude phages binding to the His-tagged epitope. Three rounds of biopanning were performed to

enrich for scFv-phages targeting AGR2 (Figure 6.4A). In each round biopanning, the phage-scFv library pool binding to the immobilised target was detected using M13-HRP monoclonal antibody in ELISA as relative light unit (RLU). The biopanning was carried out successfully, as the scFv phage pool binding to immobilised His-AGR2 was enriched in each round but not to negative control wells containing coating buffer (Figure 6.4A). The result also suggests that the scFv antibody library is bioactive that it can bind the target antigen immobilised in a microtiter plate.

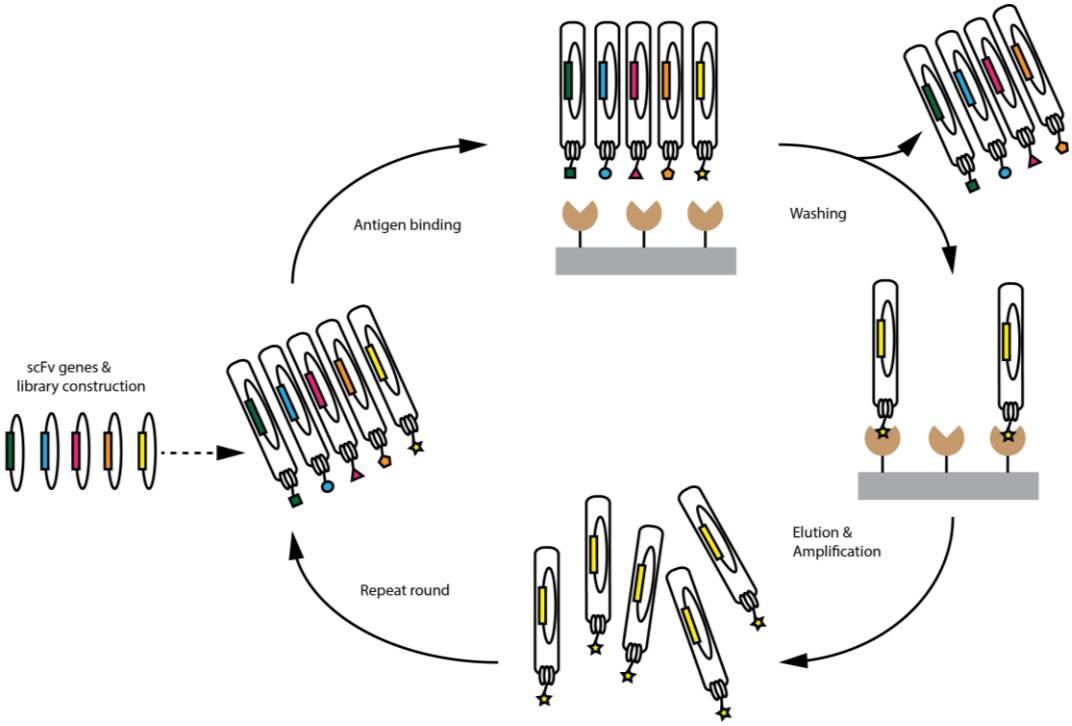


Figure 6.2 A schematic representation of the biopanning procedure in the scFv-phage display. The scFv library was made by inserting the scFv sequences into the phagemid that also encodes a recombinant phage coat protein gene III (at the C-terminus of the scFv). Infection of these transformed bacteria with helper phage enabled them to make a collection of functional phage particles expressing the scFvs fused to the surface protein pIII. A phage particle displaying a functional antibody fragment in the form of scFv as a fusion protein with the coat protein pIII. Next, the scFv-phage library is incubated with immobilised antigen coated onto a solid phase. The unbound phage antibodies are removed, and the bound ones are eluted and amplified in bacteria. These phages are then extracted and used for the next round of biopanning. Several rounds of panning are usually performed to enrich the population of scFV-phage binds to the antigen. Then, individual antibodies directed against a specific antigen are selected from the enriched library for subsequent characterization.

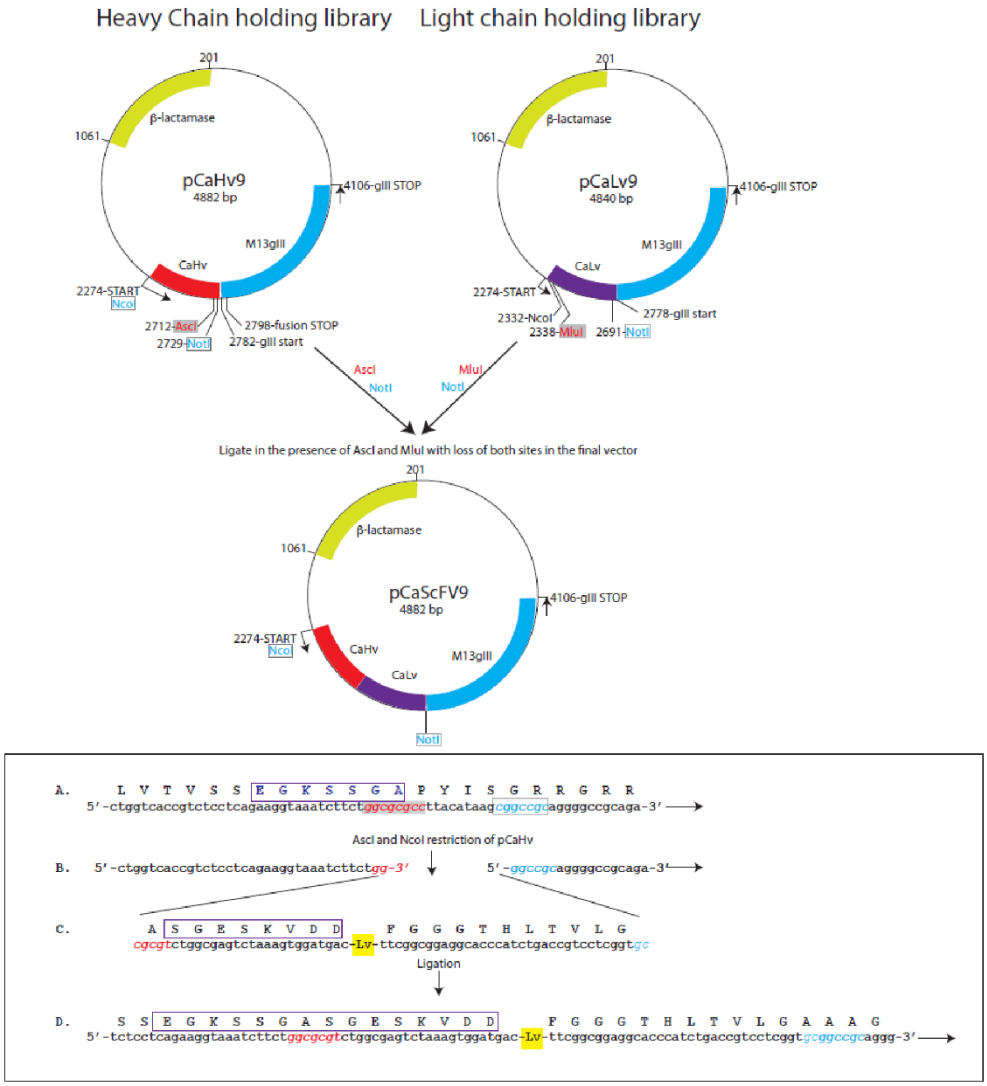


Figure 6.3 Construction of scFv antibody fragment library. The scFv fragments derived from canine naïve phage library build from dog spleen. The heavy chain (V_H) and light chain (V_L) variable fragments were PCR amplified and inserted into pCaHv9 and pCaLv9 holding phagemid vector respectively. The vector containing the heavy chain fragments digested using *AscI* and *NotI* and ligated and the vector containing the light chains were digested with *NotI* and *MluI* and ligated in the presence of *AscI* and *MluI* which were lost in the final phagemid vector, pCaScFV9. The final phagemid vector contains V_H and V_L domain DNAs joined by a 15 aa linker peptide as a single-chain Fv segment (scFv) and fused to M13 gIII coat protein. The scFv library was provided by Dr Euan Murray.

At the end of the third round of biopanning, the enriched phage-scFv library binding to AGR2 was used to infect *E. coli* TG1 cells which were plated out in LB agar plates to isolate individual phage displaying scFv antibody fragment producing plaque (monoclonals) that binds to AGR2. Forty phage producing plaques were randomly selected and grown overnight, and the scFv-phage in the supernatant was PEG precipitated. The phage-bound scFvs were screened for binding to AGR2 as quantified using the M13-HRP antibody as RLU in an ELISA. Nine high-affinity scFv were identified to bind to AGR2, and this was numbered from 1 to 9 (Figure 6.4B).

The nine high-affinity scFv binding to AGR2 were selected, grown overnight, and the scFv-phage in the supernatant was PEG precipitated. The binding of the scFv-gIII-fusion phages to a target His-AGR2 was re-assayed in an ELISA to make sure that the newly amplified phage binds to His-AGR2. The ELISA results showed that all nine phage-scFvs bound to His-AGR2 but not to negative control wells containing coating buffer (Figure 6.5A). We also tested the nine high-affinity scFv to AGR2 binding against a different His-tagged protein, His-MDM2 Ring domain (a kind gift from Dr Fiona Lickiss Prof Hupp Lab) to ensure that the phage-scFv is AGR2-specific. The PEG-precipitated phage-scFv showed binding only to AGR2 protein but not to MDM2 Ring domain (Figure 6.5B). This result also implied that the scFv did not bind to the N-terminal His-tagged epitope that is fused to the N-terminal of MDM2 Ring domain suggesting specificity to AGR2.

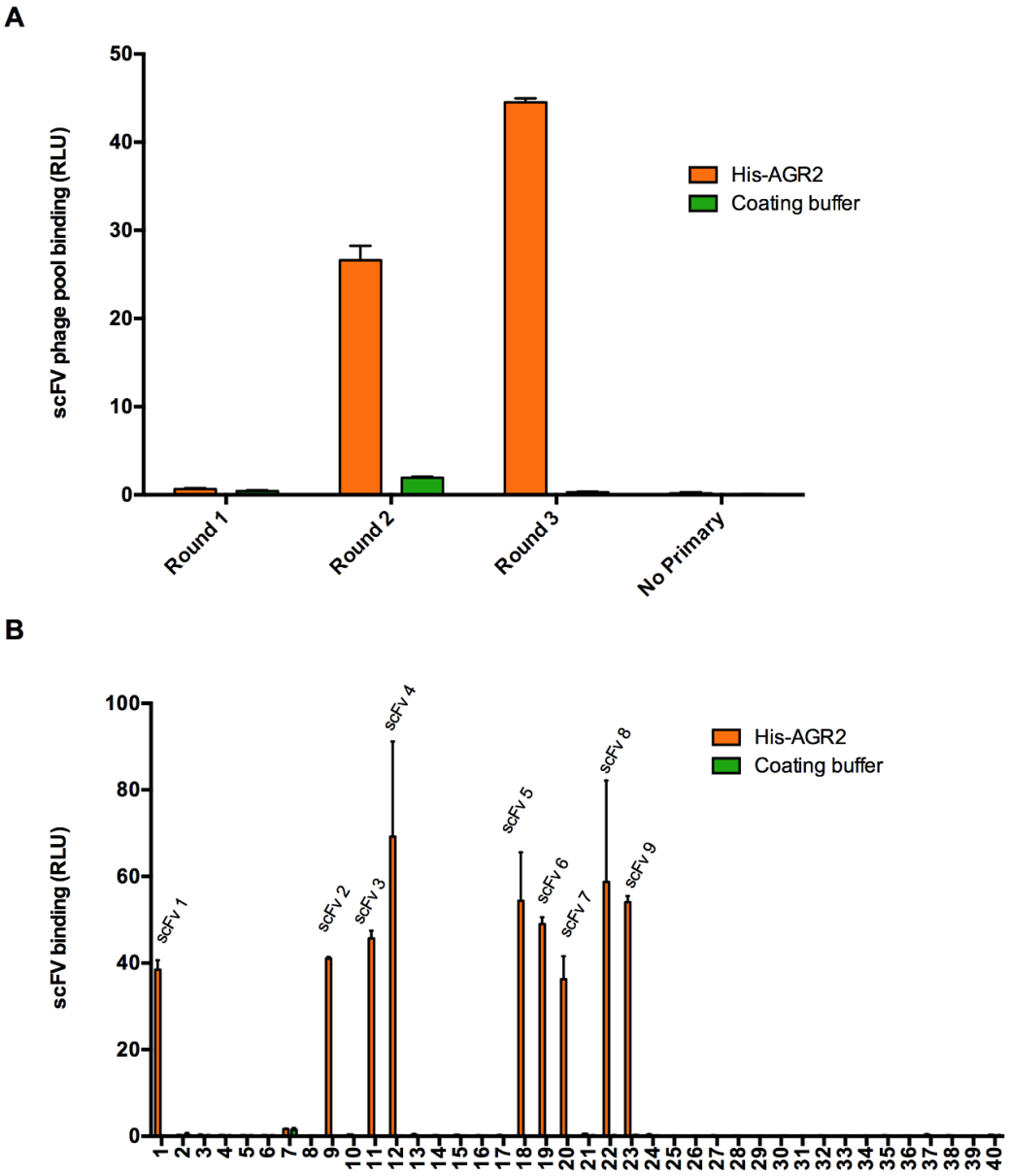


Figure 6.4 Screening of scFv-phage library binding to His-AGR2 in ELISA. (A) Enrichment of scFV-phage targeting AGR2 in a biopanning procedure. His-AGR2 (1µg in 50 µl per well) was coated onto the solid-phase of a microtiter well and used as an antigen in the selection of scFv-phage that bound to the His-AGR2. Active scFv-gIII-fusion phage binding to target antigen was detected using mouse monoclonal M13-HRP antibody as relative light units (RLU). Three rounds of panning were performed to enrich for phage library that binds to His-AGR2. The data was a representative of one experiment done in three triplicates and repeated in three independent experiments. Bars represent mean of the triplicates and and error bars show SEM. (B) Colonies (40) from rounds 3 of panning were randomly selected and grown in 200 µl of LB media and the supernatant was PEG-precipitated. The precipitated phage-bound scFV binding to immobilised His-AGR2(0.1 µg in 50 µl per well) was quantified using M13-HRP antibody as RLU. Nine high-affinity scFv-phage binding to AGR2 were identified which were numbered from 1 to 9. The experiment was performed in two independent experiments in triplicates. The data was a representative of one experiment and error bars show SEM.

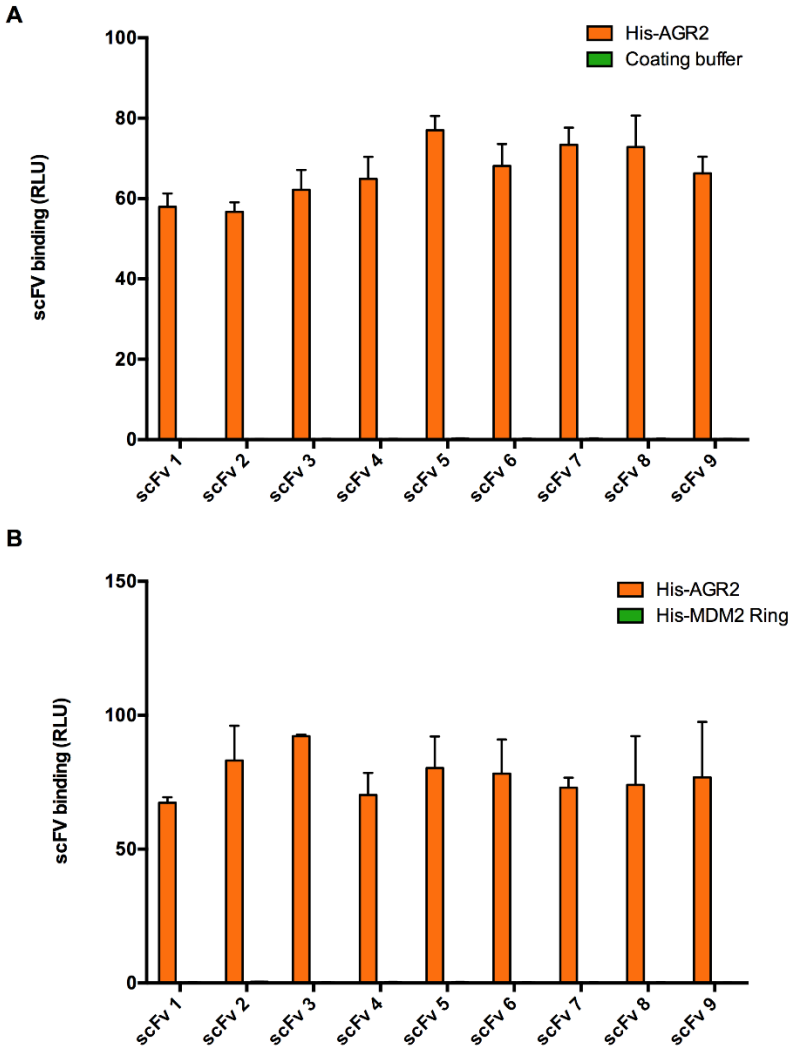


Figure 6.5 Assessing the specificity of the isolated high-affinity scFv to AGR2. (A) Phage producing plaques from nine high-affinity scFV to AGR2 were re-amplified by transfection into TG1 cells, and the scFv-phage in the supernatant was PEG precipitated, and assayed in an ELISA using His-AGR2 (0.1 µg in 50 µl per well) coated onto a solid phase. The re-amplified scFv-gIII-fusion phage was assayed using an M13-HRP phage antibody as RLU to ensure binding to AGR2 and reproducibility (B). The PEG-precipitated scFV-gIII-fusion phages were assayed for activity as in (A) using a different his-tagged protein, His-MDM2 ring domain (0.1 µg in 50 µl per well) as a negative control to ensure specificity of the scFv to AGR2 but not to His-tag or other protein. The experiment was performed in two independent experiments in triplicates. The data was a representative of one experiment and error bars show SEM.

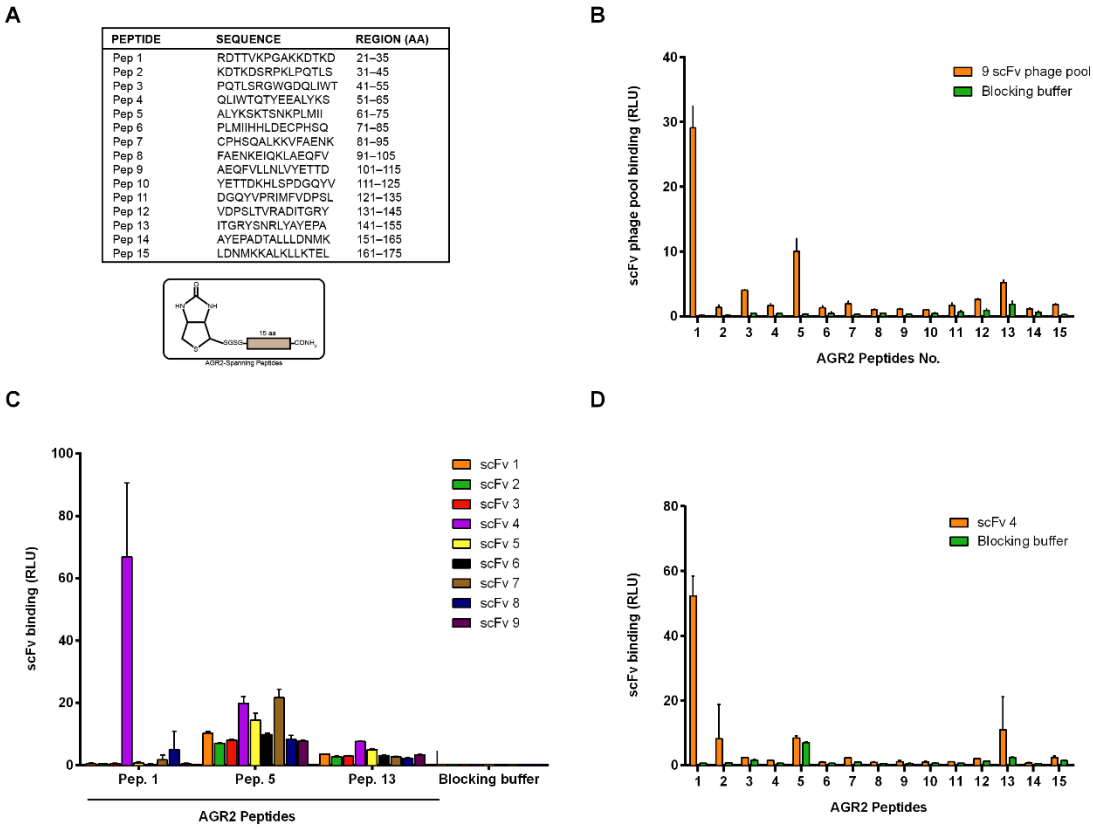
6.2.1.3. Epitope mapping of high-affinity scFv antibody fragments on AGR2 protein

We then asked which region on AGR2 does the scFv fragments bind to. In order to do so, we mapped the epitopes on individual scFv modules on AGR2 (Figure 9A-D). Identification of binding sites of an antibody is physiologically important as it can aid in the future manipulation of the antibody and may contribute to the discovery and development of new pharmacological interventions. We used an overlapping synthetic peptide library derived from an AGR2 open reading frame (ORF). The overlapping synthetic peptide library is composed of 15 amino acids with 5 amino acids overlap, and each peptide contains an N-terminal biotin-SGSG linker (Figure 6.6A). The biotinylated peptides were captured onto streptavidin coated microtiter plate, followed by incubation with the phage-scFv clones and binding was quantitated by probing with M13-HRP antibody. Incubation of the combined phage-scFv clones (polyclonals) with the biotinylated AGR2 peptide library showed that the phage-scFv pool bind with high-affinity to peptide 1, and to some extent peptide 5 and peptide 13 (Figure 6.6B). We then tested all nine individual scFv (monoclonals) binding to these three selected peptides to determine which scFv clones were responsible for binding to the peptides. ELISA results showed that one clone, namely scFv4 bound with high affinity to peptide 1 and also showed binding affinity to peptide 5 and 13 (Figure 6.6B). The other scFv clones seem to bind moderately to peptide 5 and peptide 13. It is important to note that the phage bound scFvs used were the crude lysates that obtained by PEG-precipitation, and were not normalised to the same concentration when assayed in the ELISA, so the assay may not be a true representative to measure the extent of high-affinity binding. We did not isolate scFv in its soluble form where the scFv is unfused to gIII M13 coat protein, as this scFv form is not stable in which our lab has observed previously [332] and this was true in our study as well, where our initial ELISA analysis from scFv clones isolated at the end of the third round of biopanning did not produce any detectable binding to AGR2.As such, the activity of the scFv was tested as an intact scFv bound to the phage. We then assayed high-affinity AGR2-targeting scFv4 binding to the entire AGR2 peptide library, and the result showed that the scFv4 bind to peptide 1 in high affinity and negligible binding to peptide 5 and 13 consistent with the previous ELISA results (Figure 6.6D).

6.2.1.4. Sequencing of AGR2-binding scFv antibody fragments

We sequenced all nine scFv antibody fragments to ensure that the clones are not the same and to certify that they are monoclonal scFvs, with no other contaminating scFv fragments. The primer was designed in the M13-gIII region that is fused downstream the light chain modules in the phagemid. The DNA sequencing of all nine clones demonstrated a single population of scFv indicating they are monoclonal clones but some of the scFV sequences were the same. The DNA sequences were translated using ExPasy[16] translate tool, and the deduced amino acid sequences were aligned and compared with the previously sequenced scFv phage library. This comparison helped us to determine the framework (FR) and complementary determining region (CDR) and to define areas of presumed similarity since some scFv fragments seem to bind peptide 5 of AGR2. Overall, there were five different classes of scFv antibody fragments that bind to AGR2 (Figure 6.7A). In general, the FR regions showed a high degree of homology across all scFv fragments with most of the different amino acids having been exchanged by groups with similar properties. The CDR regions in both heavy and light chain demonstrated high variations among all five classes of scFv, especially in the CDR3 heavy chain suggesting hypermutation events of both heavy and light chains during library construction. Interestingly, scFv 4 which has the highest binding affinity to purified AGR2 protein and the N-terminal AGR2 peptide is unique as it has a slight dissimilarity in the FR region and has shorter CDR3 identity implying that this region contributes to the antigen contact. The amino acids sequence comprising functional scFv4 was illustrated in Figure 6.7B which follow the direction 'heavy chain-Linker-Light chain' (V_H-L-V_L).

[16] http://web.expasy.org/translate/



PEPTIDE	SEQUENCE	REGION (AA)
Pep 1	RDTTVKPGAKKDTKD	21–35
Pep 2	KDTKDSRPKLPQTLS	31–45
Pep 3	PQTLSRGWGDQLIWT	41–55
Pep 4	QLIWTQTYEEALYKS	51–65
Pep 5	ALYKSKTSNKPLMII	61–75
Pep 6	PLMIIHHLDECPHSQ	71–85
Pep 7	CPHSQALKKVFAENK	81–95
Pep 8	FAENKEIQKLAEQFV	91–105
Pep 9	AEQFVLLNLVYETTD	101–115
Pep 10	YETTDKHLSPDGQYV	111–125
Pep 11	DGQYVPRIMFVDPSL	121–135
Pep 12	VDPSLTVRADITGRY	131–145
Pep 13	ITGRYSNRLYAYEPA	141–155
Pep 14	AYEPADTALLLDNMK	151–165
Pep 15	LDNMKKALKLLKTEL	161–175

Figure 6.6 Epitope mapping of the scFv antibody fragments binding to AGR2 protein. (A) An overlapping synthetic peptide library derived from AGR2 open reading frame. The overlapping synthetic peptide library is composed of 15 amino acids with 5 amino acids overlap, and each peptide contains an N-terminal biotin-SGSG linker (pictured in lower panel). The peptides were numbered as illustrated along with the protein sequence composition. (B) The biotinylated peptides were captured onto streptavidin coated microtiter plate. The nine high-affinity scFv-phage clones were combined, making polyclonal antibodies and the phage pool was incubated with the immobilised biotinylated peptides. Blocking buffer (3% BSA) was used as negative control. Binding was quantitated by probing with M13-HRP antibody as RLU. (C) AGR2 peptides that can be recognised by the phage pool was selected. The binding of individual nine scFv-phages was tested on these peptides (Pep 1, 5 and 13) and blocking buffer as a negative control and detected as in (B). (D) scFv4 phage which showed high-affinity to the N-terminal of AGR2 (Pep 1) was selected and assayed on all 15 AGR2 overlapping peptides to ensure epitope specificity. All data were representative of one experiment from three independent experiments done in three technical replicates. Error bars show SEM.

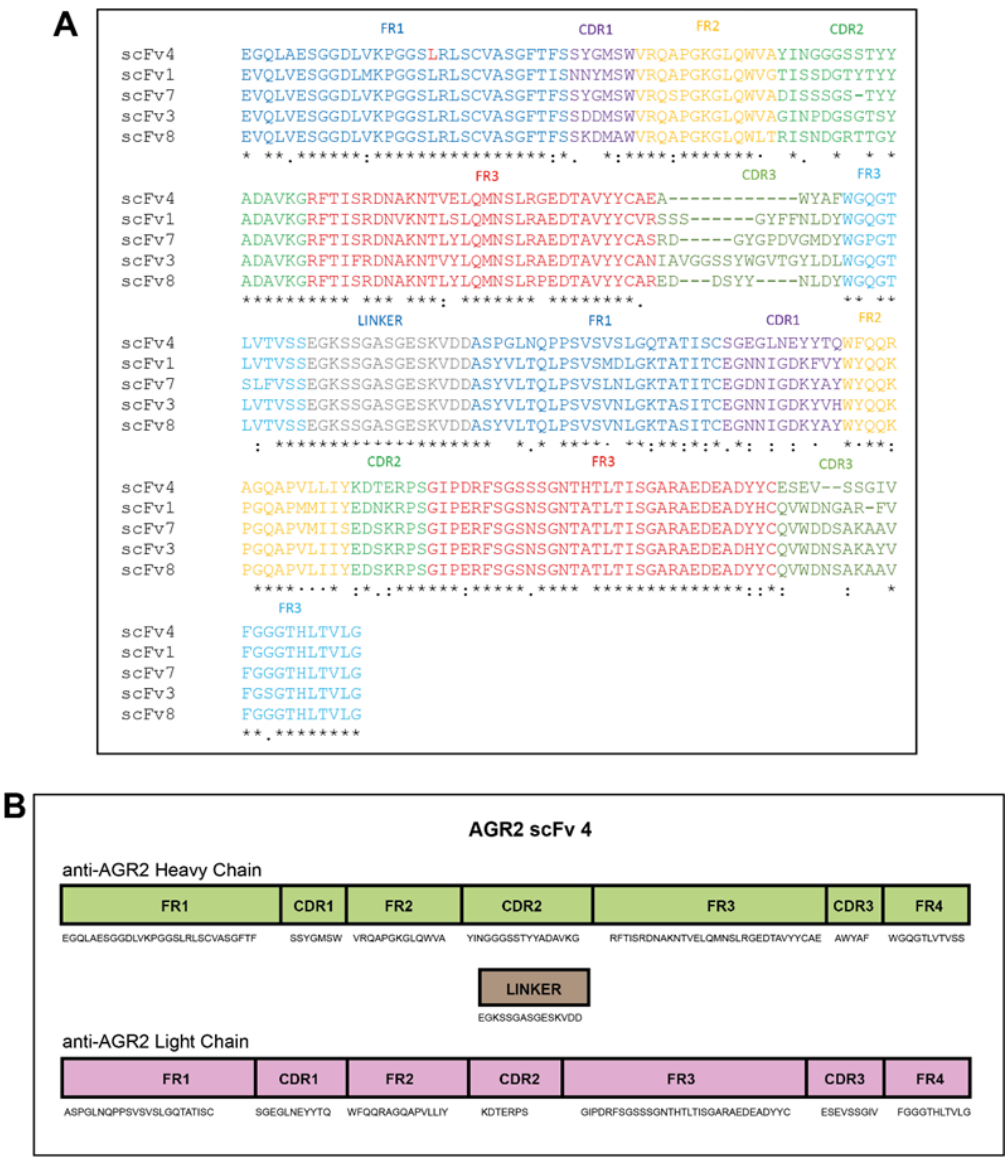


Figure 6.7 Multiple sequence alignment of the deduced amino acids obtained from DNA sequencing of AGR2-targeting scFv antibody fragments. (A) Phagemid encoding the scFv fragments were isolated and sequenced using a primer designed in the M13 gIII region. The DNA sequences were translated into protein sequence and subjected to multiple sequence alignment using Clustal Omega[17]. The amino acid sequences were compared to previously sequenced scFv heavy and light chains during library construction to determine the framework regions (FR), complementary determining regions (CDR) and linker. The scFv sequences follow the direction 'heavy chain-Linker- Light chain' (V_H - L-V_L). Different blocks of FR and CDR were coloured differently. ':' (colon) indicates conservation between groups of strongly similar properties, scoring >0.5 in the Gonnet PAM 250 matrix; '.' (period) indicates conservation between groups of weakly similar properties, scoring ≤ 0.5 in the Gonnet PAM 250 matrix. (B) Amino acids sequence of high-affinity scFv 4 targeting AGR2 highlighting FRs and CDRs of V_H and V_L.

[17] http://www.ebi.ac.uk/Tools/msa/clustalo/

6.2.1.5. Construction of monoclonal AGR2-binding scFvs into mammalian expression vector

So far, we have successfully isolated 5 different classes of scFv from the massively diverse scFv library and the scFv library contains bioactive antibodies that can detect both purified full-length protein and peptide fragments. One particular clone, namely scFv4 exhibited high-affinity binding which is epitope mapped on the N-terminal of AGR2. Significantly, the N-terminal region of AGR2 is an intrinsically disordered domain which may not interfere with AGR2 core activity, for example, the thioredoxin domain is located centrally, and the peptide binding loop is towards the C-terminal region. However, it could impact on AGR2 dimerization as deletion of the N-terminal intrinsically disordered domain can allow AGR2 to form a stable dimer [97]. We selected the high-affinity scFv4 antibody fragment to be cloned into a mammalian expression vector to see whether its expression has any impact on AGR2 or whether it can be positively or negatively by AGR2 expression. The scFv4 sequence was engineered to contain an N-terminal ER-leader sequence from AGR2 and a KDEL ER-retention site, with the control being scFv with truncated signal peptide and without the ER postcodes (Figure 6.8A). A none-AGR2-binding scFv was used as a control that also has the same AGR2 leader sequence and KDEL (ER retention motif). This scFv was derived from the screening of scFv that binds to C-terminus of HSC70-Interacting Protein (CHIP), and the isolated scFv has been validated to bind CHIP and $CHIP^{K30A}$ mutant (scFv-CHIP (clone 11F) sequences kindly provided by Professor Kathryn Ball, unpublished data). We also engineered a CD20 epitope to monitor the production of the protein. This epitope tag is a 20 amino acids peptide derived from canine CD20, that can be recognised by a mouse monoclonal CD20 antibody (NCD1.2) previously developed in our lab [332] and this was fused upstream the ER retention motif KDEL (Figure 6.8A). The three scFv constructs were human codon optimised and chemically synthesised by GeneArt synthesis (Thermo Fischer Scientific) and contain 5' *NcoI* and 3' *XbaI* restriction sites. The cDNA constructs were then subcloned into a pSF-CMV mammalian expression vector which we have extensively used in cloning AGR2 and EpCAM cDNAs in Chapter 4, where the vector is under the control of the strong CMV promoter to drive gene expression. The protein

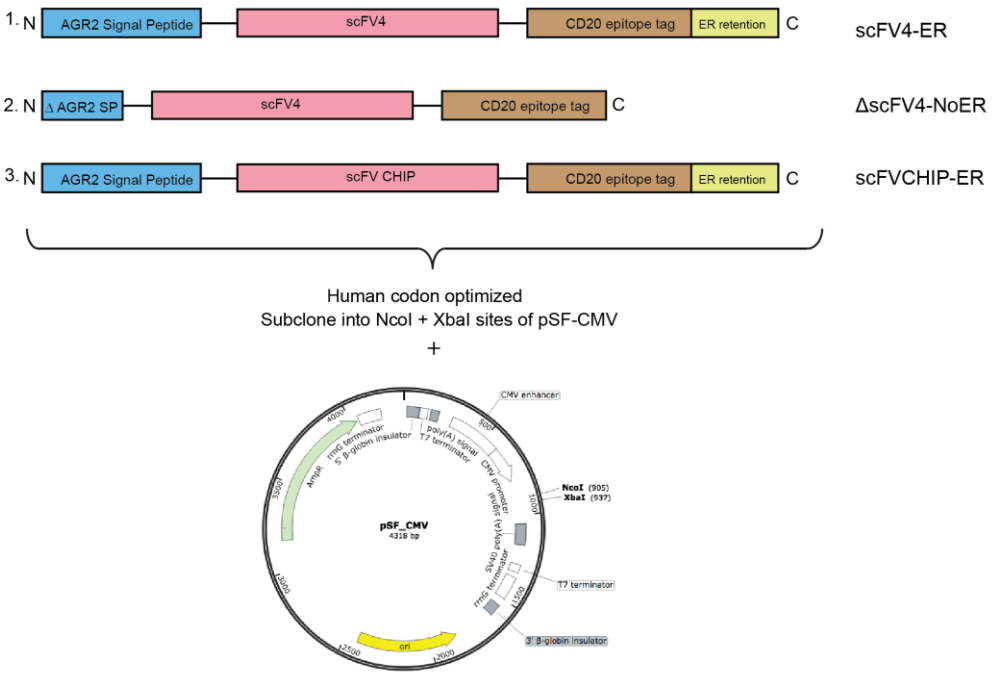
sequences of the synthetic scFv constructs using ProtParam[18], which can compute both physical and chemical parameters of a protein and the analysis is summarised in Figure 6.8B. All the synthetic scFv proteins have molecular weights between 28 kDa to 30 kDa and in general, the instability index classifies the proteins as stable.

6.2.1.6. Expression of the synthetic monoclonal scFv antibody fragments in mammalian cells

Using these vector constructs we asked whether the steady-state levels of these synthetic scFv proteins in cells can be impacted in an AGR2-dependent manner. We used two different cell lines that endogenously express AGR2 (MCF-7) and do not express AGR2 (FLO-1 cells). The transfection of the three scFv fragments resulted in a defined steady-state level of production in FLO-1 cells (AGR2-) (Figure 6.9). Both scFv4-ER and scFV4-NoER produced a single expected band at 30kDa and 28 kDa respectively as analysed by western blot. The scFv-CHIP-ER also produced a single band at 30kDa, but the expression is significantly higher suggesting that it has a more defined protein steady-state level. By contrast, in MCF7 cells (AGR2+) the scFv4-ER was substantially lower expressed and needed higher exposure for the appearance of the corresponding band in western blot (Figure 6.9). On the other hand, the expression of the scFv4-No-ER and the negative control scFv-CHIP-ER showed relatively similar expression across the two cell lines. These data suggest that the expression of the scFv4 was depended upon the ER entry/retention sequences. Intriguingly, the presence of AGR2 in MCF-7 cells can suppress the production of a high-affinity binding protein when it is destined for the ER. The scFv scaffold with an additional epitope tag and ER postcodes were able to be expressed in mammalian cells and immunoreactive as detected by western blot.

[18] http://web.expasy.org/protparam/

A



B

Plasmid names	scFV names	No. aa	MW	pI	Instability index	Aliphatic index	Hydropathicity (GRAVY)
pAM8	scFv4-ER	287	30.12	4.72	32.36	69.69	-0.264
pAM9	scFv4-NoER	269	28.17	4.76	33.86	63.46	-0.359
pAM10	scFv-CHIP-ER	295	30.86	5.42	36	71.05	-0.225

Figure 6.8 Construction of scFv antibody fragments into mammalian expression vectors. (A) The 20 aa of AGR2 signal peptide (blue) was fused to scFv4 scaffold (pink), and 20 aa CD20 epitope (chocolate) and ER retention motif KDEL (yellow) was fused at scFv4 C-terminal as illustrated in (1). As a negative control truncated N-terminal AGR2 signal peptide and deletion of KDEL was introduced to the scFv scaffold (2). A non-AGR2 binding scFv (scFv binding to CHIP protein) is constructed the same way as in (1) and serve as a second negative control (3). All constructs were human codon optimised, chemically synthesised, and sub-cloned into a high expression mammalian vector, pSF_CMV using restriction enzyme *NcoI* (5') and *XbaI* (3') which was incorporated as part of the synthetic constructs. (B) Physical and chemical parameters of the mini-proteins using ProtParam analysis.

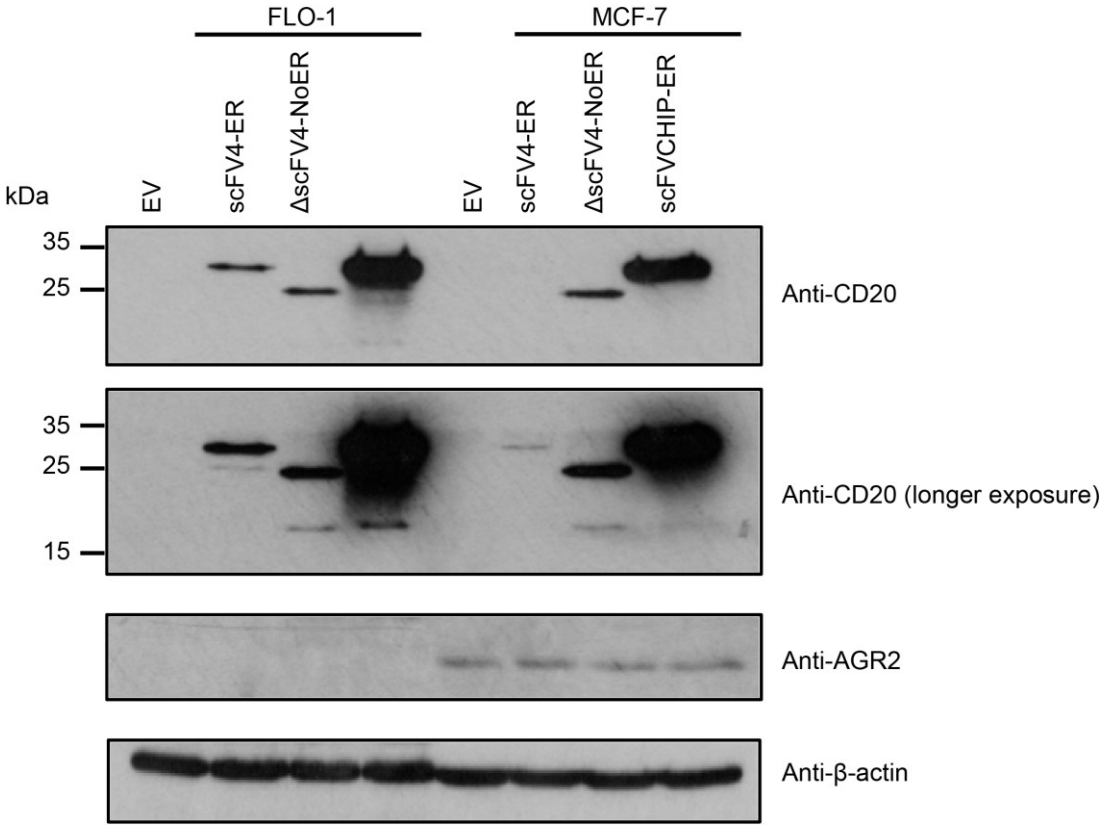


Figure 6.9 Expression of scFv modules in mammalian cells. FLO-1 or MCF-7 cells were transfected with 1 µg DNA encoding the scFV4-ER and the scFV4-NoERin a 6-well plate. As a negative control, cells were transfected with pSF-CMV empty vector (EV). Twenty-four hours post-transfection, cells were harvested and lysed using urea lysis and subjected to western blot analysis. Anti-CD20 (NCD1.2) was used to detect the presence of the scFv proteins. AGR2 was detected with anti-AGR2 K47 and anti-β -actin was used as loading control.

6.2.2. Development of synthetic mini protein containing minimal AGR2 peptide motif

6.2.2.1. Construction of synthetic mini-proteins containing AGR2 linear peptide motif repeats

Next, we aim to further manipulate the AGR2 linear peptide motif TxIYY by expressing it in mammalian cells. Our previous results using biochemical approaches and the cell-based assay analysis suggests that AGR2 function is to elevate its client protein EpCAM protein levels and this occurs via docking to the TxIYY binding site on the EpCAM receptor. We asked whether the synthetic protein expression containing TxIYY components in mammalian cells can be positively or negatively regulated by AGR2 as observed with AGR2-targeting scFv modules. We sought to deconstruct this using synthetic mini protein the same way we constructed the AGR2-targeting scFv since this construct was expressed successfully. Since the AGR2 docking motif is short, we engineered three copies of the docking motif in tandem separated by a 4 amino acids linker (Figure 6.10A) which may also increase binding affinity to AGR2. We included the original minimal hexapeptide, PTTIYY from peptide-phage display data that maintain high-affinity binding to AGR2 [209]. Like the scFv construct, the N-terminal of the docking motif modules are fused to 20 amino acids encoding AGR2 signal peptide and the C-terminal contains ER retention motif, KDEL and a CD20 epitope tag to monitor its expression. As a negative control, we designed the mutant version of the mini protein in which the linear peptide motifs were mutated to PKTIGS. In designing this, we avoided amino acid with similar property with the AGR2 consensus peptide docking motif Tx[IL][YF][YF], so that the mutant form has no binding affinity to AGR2. All the DNA encoding the mini-protein scaffolds were human codon optimised and chemically synthesised and subcloned into pSF-CMV using restriction enzyme *NcoI* and *XbaI* which were incorporated at 5' and 3' respectively of the DNA fragments. The pSF-CMV mammalian expression vector is under the control of the strong CMV promoter to drive gene expression, which we have extensively used in cloning AGR2 and EpCAM cDNAs in Chapter 4. ProtParam analysis of the mini-proteins showed that the wild-type AGR2 synthetic mini protein which we named as synPRO^{wt} has a molecular weight of 8.14 kDa and its mutant counterpart (synPRO^{mut}) has a slightly shorter molecular weight, 7.6 kDa despite

having the same amino acids length suggesting that they have different protein conformation (Figure 6.10B). Another notable difference is the theoretical pI where the synPRO^{mut} has a higher pI synPRO^{wt}. Both synthetic proteins were deemed stable as the instability index was below the suggested limit.

6.2.2.2. Expression of the synPRO^{mut} and synPRO^{wt} in mammalian cells

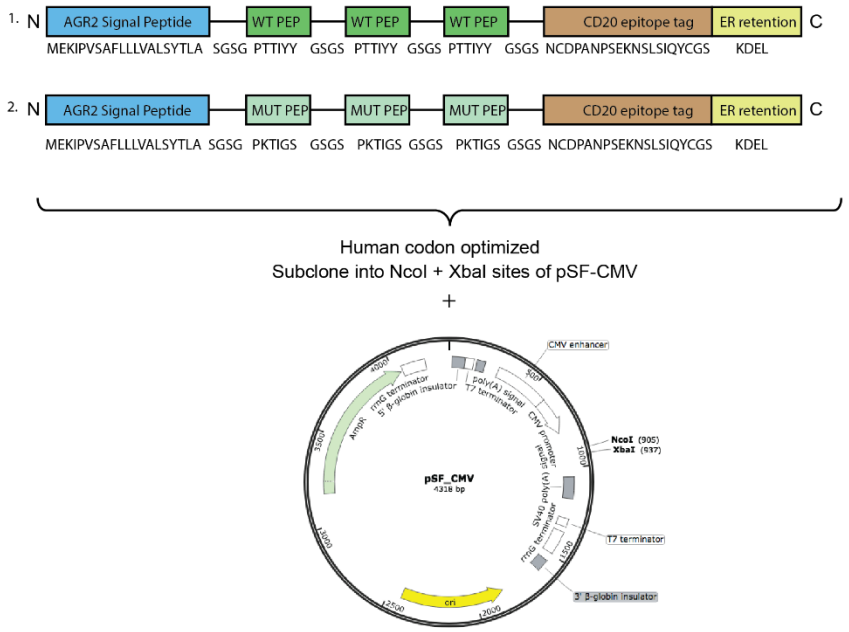
We utilised the same set of cell lines, MCF-7 and FLO-1 cells, to see if there is any difference in the expression of synthetic mini-proteins in the absence or presence of endogenous AGR2. Western blot analysis of the cells transfected with the DNA constructs encoding the mini-protein using NCD1.2 monoclonal antibody detected the expression of artificial mini protein synPRO^{mut} in both MCF-7 and FLO-1 cells (Figure 6.11A). The synPRO^{wt} can only be detected in MCF-7 cells (AGR2+) but not in FLO-1 cells (AGR2-). However, the expression of the synPRO^{wt} was substantially lower and required longer exposure for detection in western blot. The band corresponding to the synPRO^{wt} was also slightly higher than its mutant counterpart which is consistent with the ProtParam analysis. Since the synthetic mini-proteins were small in size, they may not be efficiently detected by immunoblotting which may be the case of the synPRO^{wt}. However, the detection of the synPRO^{mut} expression in both cell lines suggest that the expression of the synPRO^{wt} has lower protein steady-state level and dependent on cell types. Nevertheless, we tried to improve the detection of the synPRO^{wt} by incubating the blotted membrane after gel transfer with 0.4% PFA which was proven to increase detection of endogenous protein like α-synuclein [333]. However, the synPRO^{wt} expression is still weak compared to the synPRO^{mut} (data not shown).

We titrated the amount of DNA transfected into the MCF-7 cells to see if we can increase the expression of the synPRO^{wt}. Since the expression of the synPRO^{mut} is higher we lower the amount of DNA for transfection into the cells so that we can see a window in which the expression can be equilibrated with synPRO^{wt}. Titration of the DNA encoding the synPRO^{wt} from 0.5 µg to 4 µg showed an increased synPRO^{wt} expression at higher DNA concentration (Figure 6.11B). Lowering the amount of the DNA expressing the synPROmut (titration from 0.0625 µg to 0.5 µg) did not affect its

expression and the protein expression can still be seen at the very low level of DNA transfection. We also probed the blot for AGR2 expression to see if the endogenous AGR2 expression is affected by the expression of the mini-proteins. The result showed that the AGR2 binding peptide modules stabilise the expression of AGR2 but not in the case of the mutant binding peptide (Figure 6.11B).

Altogether, these data demonstrated that first, we were able to express the artificial mini protein in cells. The expression of the mini-protein can be seen regulated by AGR2 in cells showing different protein steady-state levels. In the case of the mini-protein containing the wild-type AGR2 binding peptide motifs, a higher amount of transfected DNA was needed for its expression to be perceived, while the mutant AGR2 peptide motif only required a small amount of DNA material for protein expression. Based on the western blot analysis, the ratio of DNA amount required for the mini-protein expression synPRO^{wt}:synPRO^{mut} is approximately 8:1. Transfection of the optimised DNA concentration for the synPRO^{wt} in FLO-1 cells (AGR2-) did not produce any corresponding band in western blot (data not shown) which further implied that synPRO^{wt} expression dependent on AGR2. The stabilisation of AGR2 when the wild-type mini-protein is expressed (but not the mutant mini-proteins) suggests that they interact with AGR2 in cells presumably through the peptide docking interface (aa 131-135) on AGR2 which we have mapped earlier (Chapter 3) using hydrogen-deuterium exchange mass spectrometry.

A



B

Plasmid names	Mini protein names	No. aa	MW	pI	Instability index	Aliphatic index	Hydropathicity (GRAVY)
pAM11	synPROwt	78	8.14	4.51	30.71	72.56	-0.109
pAM12	synPROmut	78	7.68	7.75	29.23	72.56	-0.178

Figure 6.10 Construction of mini-protein containing AGR2 binding peptide repeats into a mammalian expression vector. (A) The minimal AGR2 binding peptide (PTTIYY, green) was synthetically constructed in tandem (3x PTTIYY) separated with GSGS linker which is fused to AGR2 signal peptide at N-terminal (blue) and CD20 epitope tag (light chocolate) and endoplasmic reticulum (ER) retention motif at C-terminal (yellow) to direct the scFv to endoplasmic reticulum where AGR2 normally resides. As a negative control, a mutant version of the AGR2 binding peptide (PKTIGS, light green) was synthesised the same way as the AGR2 'WT' binding peptide counterpart. All constructs were human codon optimised, chemically synthesised, and sub-cloned into a high expression mammalian vector, pSF_CMV using restriction enzyme NcoI (5') and XbaI (3') which was incorporated as part of the synthetic constructs. (B) Physical and chemical parameters of the mini-proteins using ProtParam analysis.

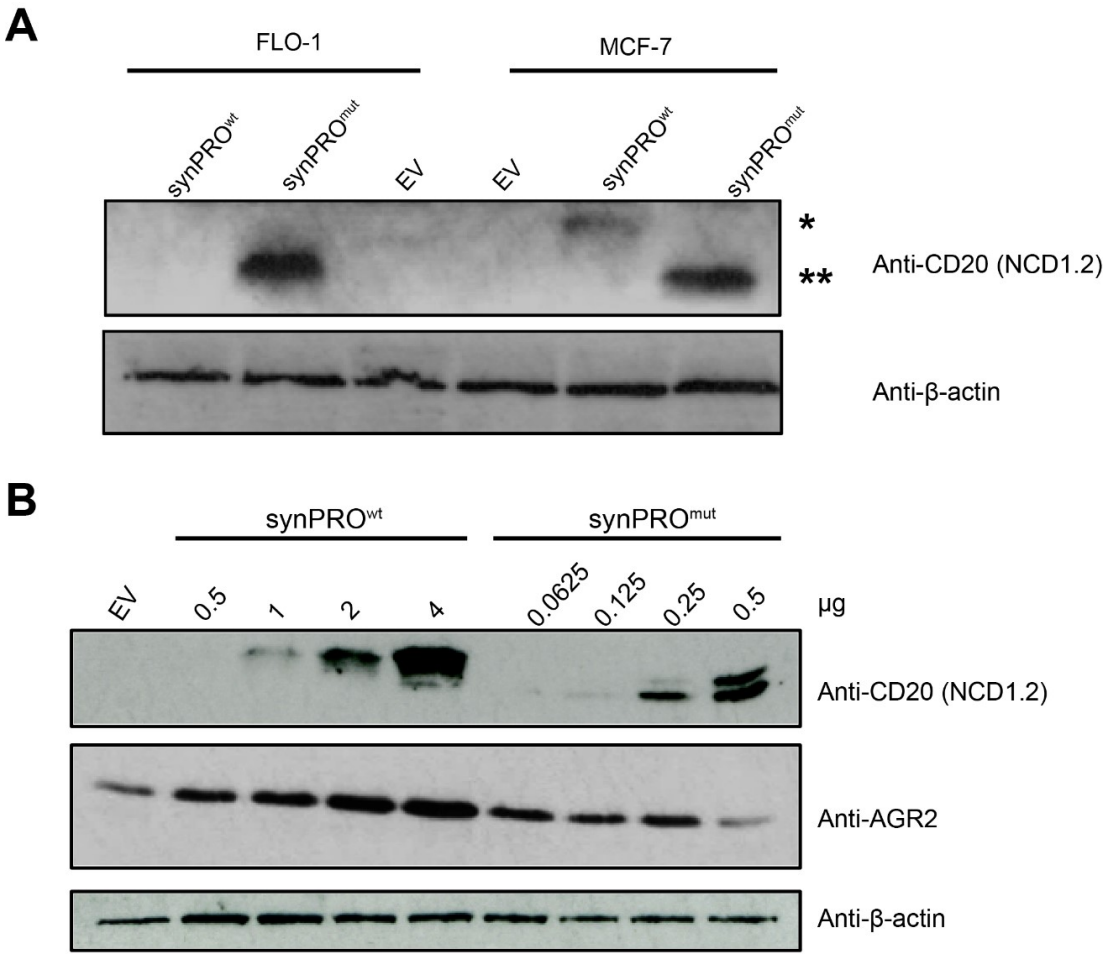


Figure 6.11 Expression of synthetic mini-protein encoding AGR2 binding peptide repeats and its mutant form in mammalian cells. (A) FLO-1 or MCF-7 cells were transfected with 1 µg DNA encoding the AGR2 binding peptides synPRO^{wt} and the mutant form synPRO^{mut} in a 6-well plate. As a negative control, cells were transfected with pSF-CMV empty vector (EV). Twenty-four hours post-transfection, cells were harvested and lysed using urea lysis and subjected to western blot analysis. Anti-CD20 was used to detect the presence of the mini-proteins. Expression of synPRO^{wt} is marked by (*) and synPRO^{mut} (**). β-actin antibody was used as loading control. (B) MCF-7 cells were transfected with titration of indicated DNA encoding the synPRO^{wt} and synPRO^{mut} proteins in a 6-well plate. The amount of transfected DNA is normalised with pSF-CMV empty vector. Twenty-four hours post-transfection, cells were harvested and lysed using urea lysis and subjected to western blot analysis. Anti-CD20 (NCD1.2) was used to detect the presence of the mini-proteins and also probed for AGR2 using anti-AGR2 (K47). β -actin antibody was used as loading control.

6.2.2.3. AGR2 and synthetic mini protein interaction using in situ proximity ligation assay

We evaluated whether the mini-protein containing AGR2 peptide motif repeats can form a complex in cells using proximity ligation assay (PLA) whose interaction has been established previously using *in-vitro* biochemical assays. Since the mini-protein containing the wild-type AGR2 peptide repeats only express in the cells that endogenously express AGR2, we utilised MCF-7 cells as a model for the PLA assay. Based on our prior result, transfection of the same amount of DNA give different expression of the wild-type and mutant mini-proteins levels. Therefore, we transfected the cells according to the optimised DNA amount that achieves the same protein expression level of the wild-type and mutants mini-proteins so we can assay the protein-protein interaction accurately (Figure 6.11B). We utilised the Duolink In-situ based PLA assay that we have used before to detect the AGR2 and EpCAM protein-protein interaction (Chapter 4). This assay requires two primary antibodies each directed against one of the targets of interest and the two primary antibodies must be raised in different species. We used the mouse monoclonal CD20 (NCD1.2) for detecting the transfected mini-proteins and rabbit polyclonal AGR2 (K47) directed against the endogenous AGR2 in MCF-7 cells. PLA assay incubated with the antibody pair showed PLA signals in cells transfected with synPRO^{wt} (Figure 6.12, upper panel). This suggests that the mini-protein containing repeats of AGR2 binding peptide form complex with endogenous AGR2 indicating protein-protein interaction in cells consistent with our previous *in-vitro* biochemical assay. On the other hand, the mutant form of the mini-protein, synPRO^{mut} did not produce PLA foci indicating that there is no protein-protein interaction taking place in the cells Figure 6.12, middle panel). As a control, MCF-7 cells were transfected with pSF-CMV empty vector (EV) and incubated with the antibody pair in PLA which exhibited no foci (Figure 6.12, lower panel).

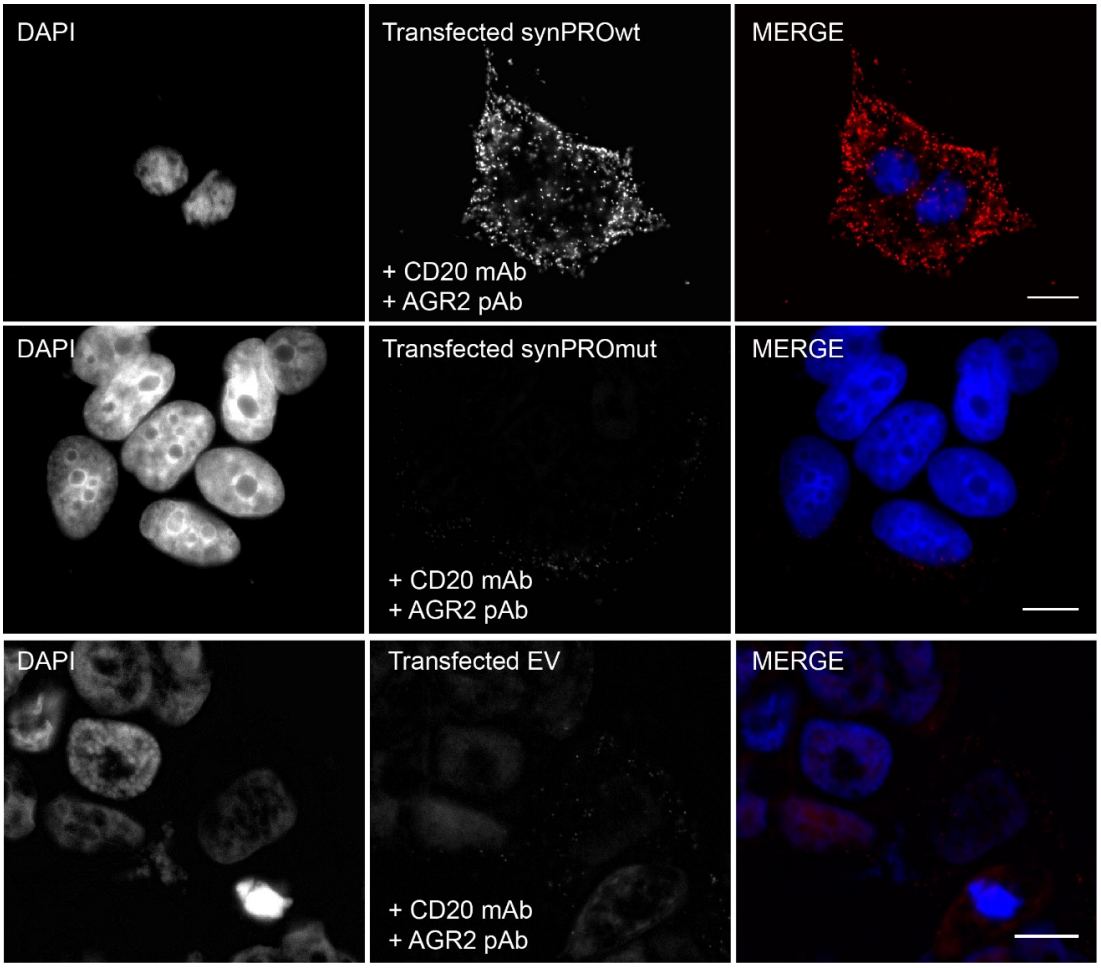


Figure 6.12 AGR2 and synthetic mini-proteins interaction using in situ proximity ligation assay. (A) MCF-7 cells were transfected with DNA encoding synPRO^{wt} (4µg, upper panel); synPRO^{mut} (0.5 µg), or pSF-CMV empty vector (2 µg). These DNA concentrations were according to optimised starting DNAs to give equal protein expression of synPRO^{wt} and synPRO^{mut} .After 24 hours transfection, the cells were processed by Duolink In situ proximity ligation assay using mouse monoclonal CD20 (NCD1.2) and rabbit polyclonal AGR2 (K47) antibody pair. Following ligation and amplification, protein-protein interaction complex was detected with Duolink red fluorescent probes. Red fluorescent foci indicate the interaction between the two proteins. Nuclei were counterstained with DAPI and cells were visualised with an epifluorescence microscope. Scale bar 10 µm.

6.2.3. Development of synthetic membrane protein containing AGR2 peptide motif repeats

6.2.3.1. Designing and construction of synthetic membrane protein containing AGR2 peptide motif repeats

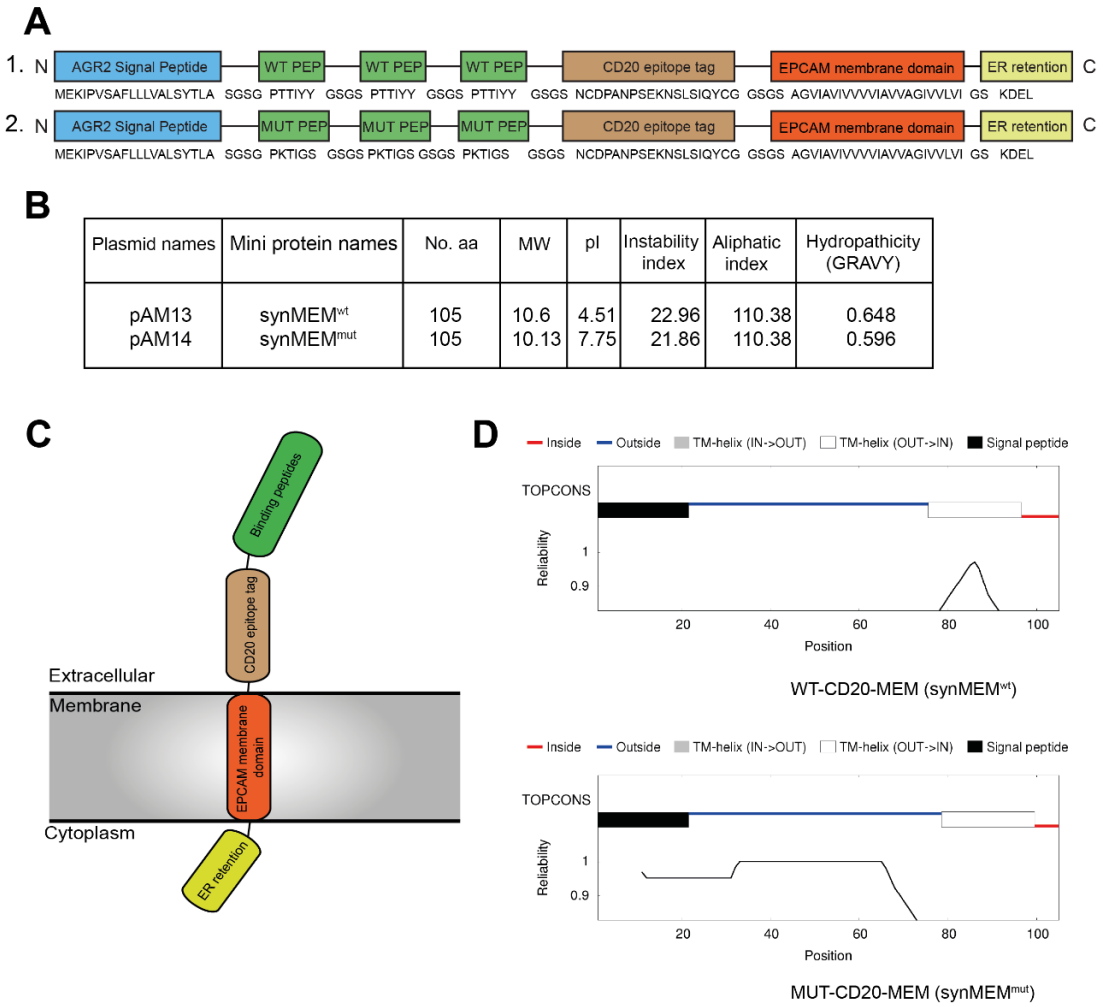
Our results so far suggest that AGR2 can bind synthetic mini-protein containing its linear peptide motif in cells, complementing our *in-vitro* protein-protein interaction assay. AGR2 was shown to elevate its client protein level containing such motif, EpCAM which occurs via docking to the TxIYY binding site on the EpCAM receptor. We also showed that the delivery of EpCAM to the plasma membrane is mediated by AGR2. Therefore, we wanted to create an artificial membrane protein containing a transmembrane domain and AGR2 binding peptide motif to see whether the artificial membrane exhibits the same activity as the EpCAM membrane protein. The synthetic membrane protein consists of the same composition of the mini-protein containing AGR2 binding peptide repeats but with an addition of transmembrane domain (Figure 6.13A). This domain comprises of 23 amino acids that are derived from the EpCAM transmembrane helix (amino acid 266-288). We added the transmembrane domain downstream of the epitope tag upstream in which we predict that the epitope tag will appear in an extracellular orientation which as illustrated in Figure 6.13C. Our prediction is based on the EpCAM structure where the TLIYY motif is located in the extracellular stalk. This orientation allows us to visualise plasma membrane localisation using CD20 antibody recognising the extracellular side in immunofluorescence study. We also analysed the synthetic membrane protein sequences using TOPCONS[19] which can predict membrane protein topology and can efficiently separate signal peptides from transmembrane regions [334]. The analysis results were in agreement with our membrane topology models which suggest that the epitope tag is on the extracellular side (Figure 6.13D). However, it is also important not to rule out that there is a possibility that the orientation can be in the reversed orientation. ProtParam analysis of the proteins suggests that the proteins are more hydrophobic (positive value of hydrophaticity) which is consistent with the addition of the transmembrane domain (Figure 6.13B). The synthetic membrane protein containing the AGR2 binding peptide repeats or the mutant form was human codon

[19] http://topcons.net/

optimised and chemically synthesised. These were subcloned the same way as the AGR2 mini-proteins into pSF-CMV via restriction site *NcoI* (5') and *XbaI* (3').

6.2.3.2. Expression of the synthetic membrane proteins

We utilised the same set of cell models MCF-7 (AGR+) and FLO-1 (AGR2-) for the expression of the synthetic mini-proteins. Since we have shown that synthetic mini-proteins without the transmembrane domain are dependent on endogenous expression of AGR2, we ask whether the synthetic membrane proteins have the same activity and whether the presence of AGR2 can influence the delivery of the synthetic membrane protein to the plasma membrane as in the case of EpCAM, since the synthetic proteins have EpCAM's transmembrane domain. Western blot analysis of the MCF-7 (AGR2+) cells transfected with equal amount DNA constructs encoding the synthetic wild-type or mutant membrane proteins showed different protein steady-state levels (Figure 6.14A). The synMEMwt is substantially lower compared to the synMEMmut. Transfection into FLO-1 (AGR2-) showed much lower expression of both membrane protein (Figure 6.14B). These data suggest that the expression of the membrane proteins were regulated by AGR2 consistent with the expression of the synthetic mini-protein (synPRO) we have previously observed. The lower expression of the wild-type membrane protein may indicate that it interacts with AGR2 via docking to the PTTIYY for proper maturation and trafficking to the plasma membrane or possibly secreted into media, hence lower total protein load upon harvesting for western blot. The mutant, however, cannot bind to the AGR2 which may be predominant in soluble form in the cytosol, therefore, more abundance in the cells when harvested for western blot analysis.



Plasmid names	Mini protein names	No. aa	MW	pI	Instability index	Aliphatic index	Hydropathicity (GRAVY)
pAM13	synMEM^{wt}	105	10.6	4.51	22.96	110.38	0.648
pAM14	synMEM^{mut}	105	10.13	7.75	21.86	110.38	0.596

Figure 6.13 Construction of synthetic membrane proteins containing AGR2 binding peptide repeats and EpCAM transmembrane domain. (A) The synthetic membrane proteins contain the same structure of the synthetic mini-protein (synPRO) but with an addition of EpCAM transmembrane domain (orange). This is inserted downstream the CD20 epitope tag which is flanked by a small stretch of linker. All constructs were human codon optimised, chemically synthesised, and sub-cloned into a high expression mammalian vector, pSF_CMV using restriction enzyme *NcoI* (5') and *XbaI* (3') which was incorporated as part of the synthetic constructs. (B) Physical and chemical parameters of the synthetic membrane proteins using ProtParam analysis. The wild-type synthetic membrane protein was named synMEM^{wt} and the mutant counterpart. (C) Schematic representation of membrane topology model based on EpCAM structure, where the transmembrane domain embedded in the plasma membrane, while the epitope tag and the AGR2 binding motifs are in the extracellular stalk orientation. (D) Membrane protein topology TOPCONS prediction analysis which supports our model topology in (C) for synMEM^{wt} (upper panel) and synMEM^{mut} (lower panel) suggesting the epitope tag and the binding motifs are extracellular and recognise EpCAM transmembrane domain as a helix.

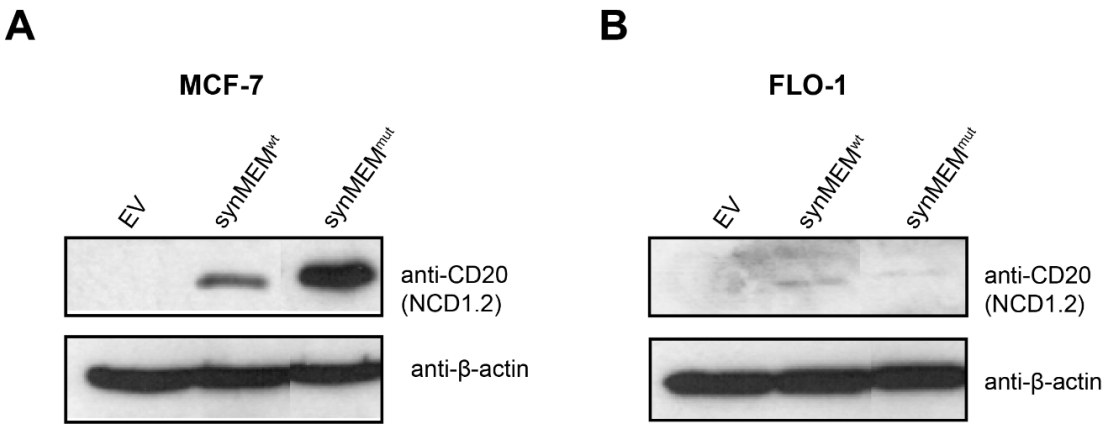


Figure 6.14 Expression of synthetic membrane protein containing EpCAM transmembrane helix and AGR2 binding peptide repeats. MCF-7 (A) or FLO-1 (B) were transfected with 1 µg od DNA encoding synMEM^{wt}, synMEM^{mut} or pSF-CMV empty vector (EV) in a 6-well plate. Twenty-four hours post-transfection, cells were harvested and lysed using urea lysis and subjected to western blot analysis. Anti-CD20 was used to detect the presence of the wild-type or mutant synthetic membrane proteins. β-actin antibody was used as loading control.

6.2.3.3. Localisation analysis of the synthetic membrane proteins using immunofluorescence

Next, we examined the localisation of the synthetic membrane protein in the cells using immunofluorescence. We ask whether if there are different localisations of the wild-type and mutant membrane proteins, in which the expressions are regulated by AGR2. We hypothesised that in the presence of AGR2, the wild-type synthetic membrane protein synMEM^{wt} is predominantly localised in the plasma membrane while the synPRO^{mut} is more abundance in the cytosol. MCF7 or FLO-1 cells lines transfected with the synthetic membrane proteins were stained with mouse monoclonal CD20 (NCD1.20) recognising the epitope tag predicted to be at the extracellular side. Cells were then fixed and permeabilised and visualised using epifluorescence microscopy. Immunofluorescence staining of the synMEM^{wt} demonstrated a mixture of localisations in MCF-7 cells, where there was a pool of plasma membrane localisation and also a cytosolic distribution appearing as puncta (Figure 6.15A, upper panel). On the other hand, the synMEM^{mut} localisation was largely cytosolic Figure 6.15A, middle panel). Interestingly, the synMEM^{mut} appeared as puncta and in some cells seemed more granular/vesicle-like distribution. In FLO-1 cells, both the wild-type and the mutant synthetic membrane proteins exhibited cytosolic distribution (Figure 6.15B, upper and middle panel). However, the wild-type synMEM distribution was more 'spotty' or as puncta, while the mutant synMEM localisation looks more diffused. As negative controls, there were no positive staining in cells transfected with CMV empty vector only (Figure 6.15A, and B, lower panel).

Overall, we were able to express the artificial membrane protein containing AGR2 binding peptide motifs, synMEM^{wt} and remarkably the synthetic protein was seen to localise partially in the plasma membrane in MCF-7 cells (AGR2+). It is important to point out that we engineered a 'strong' canonical ER retention motif KDEL in the synthetic membrane protein (Figure 6.13). This suggests that some proteins may be retained in the ER which explained why we observed some puncta staining for synMEM^{wt}. Nevertheless, the synMEM^{wt} plasma membrane distribution suggests that the protein can bind to endogenous AGR2 in the ER via docking to the PTTIYY repeats which drive the protein to enter secretory pathway or to escape ER

and finally to the plasma membrane. There was no plasma membrane localisation seen in FLO-1 cells (AGR2-) for synPRO^{wt} suggesting that the protein was unable to interact with AGR2 for its protein maturation, which may be trapped in the ER or subjected to proteasomal degradation. The synPRO^{mut} controls exhibited similar distribution where they are largely cytosolic in both MCF-7 and FLO-1 cells. However, there was a notable difference in the pattern of staining where synPRO^{mut} demonstrated more puncta/granular staining in MCF-7 cells, whilst in FLO-1 displayed diffused staining suggesting that the proteins have different regulations in different cells. Importantly, the mutant membrane protein was not delivered to plasma membrane indicating that AGR2 was not able to recognise the mutant peptide motif for docking and processed for its maturation through the secretory pathway.

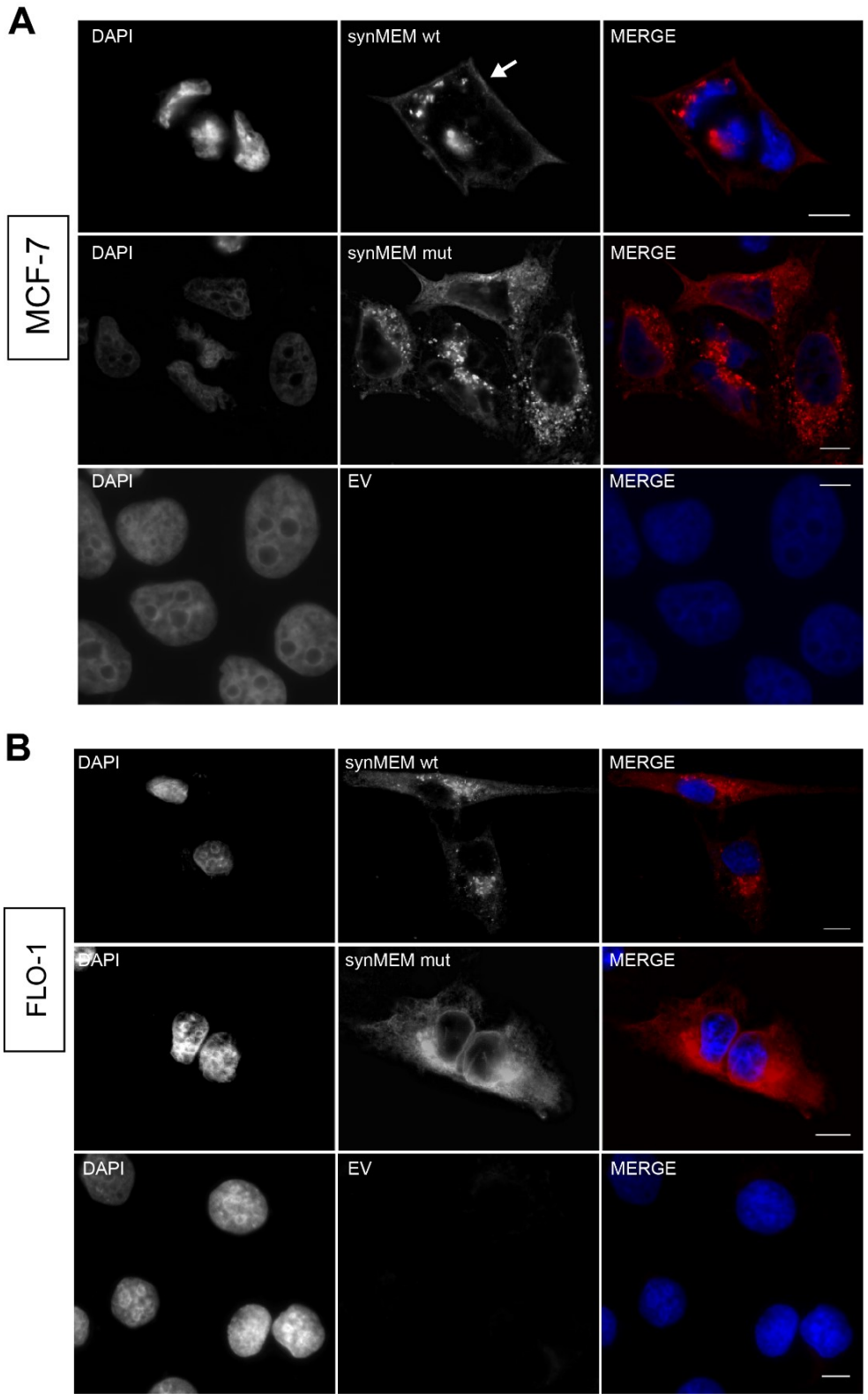


Figure 6.15 Localisation of synthetic membrane proteins containing AGR2 binding peptide motif using immunofluorescence analysis. (A) MCF-7 cells were transfected with DNA encoding synMEM^{wt} (upper panel); synMEM^{mut} (middle panel) or; CMV empty vector (EV) (lower panel) and incubated for 24 hours. The cells were then immunostained with CD20 monoclonal antibody (NCD1.2)

with a combination of Alexa 594 (red) secondary antibodies and detected using epifluorescence microscopy. Scale bar 10 μm. (B) FLO-1 cells were transfected with DNA encoding synMEM^{wt} (upper panel); synMEM^{mut} (middle panel) or; CMV empty vector (EV) (lower panel) and incubated for 24 hours. The cells were then immunostained with CD20 monoclonal antibody (NCD1.2) with a combination of Alexa 594 (red) secondary antibodies and detected using epifluorescence microscopy. Scale bar 10 μm. Nuclei were counterstained with DAPI.

6.3. Discussion

It has been more than 30 years since the first monoclonal antibodies were generated. Monoclonal antibodies (mAbs) were first generated by Kohler and Milstein in 1975 using hybridoma technique and this founding work has embraced the area of clinical research and has been acknowledged as the magic bullet against tumour antigens as well as several other diseases because of their inherent capacity for specificity [335]. The monoclonal antibody industry has expanded exponentially ever since and is the fastest growing component used in immunotherapy identifying targets in wide-range of diseases and conditions [302]. With major advances in molecular biology and biomedical research, much research into monoclonal antibodies now focuses on identifying new targets for development and maximising their efficacy for use in clinical practice. Currently, there are approximately 30 mAbs that have been approved for use in clinical practice while much more waiting in the pipeline and being tested in clinical trials.

mAbs have been used extensively in cancer therapeutics which target a wide variety of antigens expressed on the surface of cancer cells [8]. There are different mechanisms of actions of mAbs targeting cancer cells and these dependent on the mAbs construct and the nature of tumour malignancy. Some tumours specific mAbs have a direct effect on tumour cells upon binding cell surface antigen which induces apoptosis by direct transmembrane signalling [336]. There was also evidence that mAbs involved in immune-mediated effects which are called complement-mediated cytotoxicity (CMC) which directly result in target cell death through the development of a membrane attack complex (MAC) [337]. Another immune-mediated effect upon mAbs antigen contact is by inducing antibody-dependent cellular cytotoxicity (ADCC) [338] which result in enhancing phagocytosis and lysis by monocytes and granulocytes.

A key criterion to be considered for development of antibody therapeutics is to identify a suitable target in the disease and to reduce immunogenicity [339]. Other limitations include side-effects, inefficient models for mAb generation and a lack of efficacy. Different manipulative forms of antibodies have been developed such as

humanised, chimeric and fully human antibodies which aim to overcome the issues such as human anti-mouse antibody response raised by mouse monoclonal antibodies [25].

As discussed previously, AGR2 has emerged to be a potential drug target in cancer. It was found to be overexpressed in various type of cancers [94, 116]. AGR2 showed pro-oncogenic features, where its overexpression, for example, can enhance cancer cell survival, rather than inhibit cell growth in the premalignant Barrett oesophagus and oesophageal cancer models [87]. In addition, high levels of AGR2 are associated with poor prognosis in ER-positive breast cancer and was shown to promote metastasis of breast epithelial cells in an *in vivo* metastasis assay [120, 139], suggesting AGR2 as a potential drug target and biomarker. The clinical relevance of AGR2 is strengthened by the finding that AGR2 showed resistance to major cancer drug therapies such as Tamoxifen [171].

In this chapter, we aim to develop monoclonal antibodies targeting AGR2 in the format of scFv antibody fragments, in which we screened from a massive scFv-phage display library which roughly contains 10^9 scFv sequences. Antibody fragment in the format of scFv has been successfully modified in several applications, for example, scFv fragments were fused with radioisotopes for cancer imaging [340], scFv conjugated with enzymes in antibody-directed enzyme prodrug therapy (ADEPT) [341], scFv as immunotoxins for targeted tumour cell targeting [342] and biosensors for real-time detection of target molecules [343]. The phage library used here was constructed by in-house cloning of massive collection of scFv genes derived from the canine naïve library and cloned into phage display vectors to determine whether single chain antibody domain fusions can be packaged in bioactive form (Figure 6.3). We were able to isolate five classes of high-affinity scFv antibody fragments that bind to specifically to purified His-AGR2 and this as well suggests that the scFv antibody library is bioactive that it can bind the target antigen (Figure 6.4 and Figure 6.7). Epitope mapping using 15-mer overlapping peptides of AGR2 ORF demonstrated that all five scFvs have an affinity towards centrally located peptide no 5 (aa 61-75) and to some extent to peptide 13 (aa 141-155) (Figure 6.6B-CFigure 6.8). One particular scFv

however, showed high-affinity binding to peptide 1 which is the N-terminal intrinsically disordered region of AGR2 (aa 21-35). This epitope is away from the core domain of AGR2 such as the thioredoxin domain CPHS (aa 81-84), peptide binding loop (aa 104-111) or peptide docking site we have mapped in Chapter 3 (aa 131-135) but however may affect AGR2 dimerisation [97]. We selected this scFv4 and cloned it into a mammalian expression vector which was engineered to contain ER postcodes for entry into ER where AGR2 normally resides and CD20 epitope to monitor its expression (Figure 6.8). We were able to express the scFv scaffolds in mammalian cells and were immunoreactive as they can be detected with a CD20 antibody (Figure 6.9). Our results showed that the expression of the scFv with ER postcodes (scFV4-ER) was regulated by the presence of AGR2 and the expression of the ER-directed scFv4 depended upon the ER entry/retention sequences, where the scFv without ER postcode showed the same expression levels in MCF-7 and FLO-1 cells. This suggests that the production of high-affinity scFv targeting AGR2 can be suppressed by AGR2 when destined for the ER. However, we have not yet checked whether the proteolytic activity is involved in regulating the scFv expression which can be done by the treatment of proteasomal inhibitor MG132 in the cells transfected with these scFv constructs. Strikingly, the expression of AGR2 remains unchanged when the AGR2-targeting scFv is expressed. One explanation is that scFv expression does not effect on the total protein load of AGR2, but may affect AGR2 function and signalling (such as enhancing the unfolded protein response or relieving ER stress) and AGR2 cellular distribution (since AGR2 can escape ER and can be secreted). There is also a probability that there might be a negative feedback loop maintaining its expression, which could be investigated in the future. This finding also expands current monoclonal AGR2 antibodies since there are only a few AGR2 mAbs available commercially. Our group has been trying to generate in-house monoclonal AGR2 antibodies using hybridoma techniques from the purified AGR2 protein as antigen were nor fruitful. Some of the isolated monoclonal antibodies showed cross-reactivity with its homologue AGR3 and cannot be used in immunoprecipitation. We are planning to express and purify these monoclonal antibodies in the future and test their potential and specificity in assays like immunoprecipitation. It is also interesting to see the effect of these scFvs expression on AGR2 client proteins in the future. In addition,

these scFvs can be translated into immunotherapy use for example as a recombinant bioconjugate system for ADEPT.

Since we were able to express functional high-affinity scFv that binds to AGR2 in cells, we ask if we can repeat the experiment by exploiting the AGR2 linear peptide motif previously isolated from peptide phage-display and construct it the same way as the synthetic scFv modules but with three copies of AGR2 linear peptide motif PTTIYY motif in tandem (Figure 6.10). The expression of the synthetic mini-protein containing AGR2 linear motifs (synPRO^{wt}) was also regulated by AGR2, where the production of synPROwt was only expressed in AGR2-expressing cells (MCF-7) but not AGR2-null cells (FLO-1) (Figure 6.11). Compared to the mutant form of mini-protein (synPRO^{mut}), the expression of synPRO^{wt} requires a higher amount of starting DNA to achieve steady-state level. Importantly, the AGR2-binding protein synPRO^{wt} stabilises AGR2 expression suggesting that they interact with each other. However, we showed using western blot only that titration of synPRO^{wt} but not its mutant counterpart stabilises AGR2. This data could be further strengthened by looking at the stability of the synPRO^{wt} and it mutant counterpart in MCF-7 and FLO-1 cells by using pulse-chase experiment and to check whether proteasomal degradation causes the lower expression of the synPRO^{wt}. We showed using PLA that AGR2 can interact with the wild-type synthetic protein but not the mutants which complement our *in-vitro* protein-protein interaction assays in Chapter 3 (Figure 6.12). This data highlight the importance of peptide-protein interactions that regulate the magnitude of signal amplification in a signal transduction pathway which may fundamentally serve as regulatory switches. We also provided evidence using a cell-based system which may be further tested in animal models for potential peptide lead in the development of peptide therapeutics to disrupt pro-oncogenic AGR2 protein-protein interactions in cell systems. Such peptide therapeutics (peptide-mimetics) have been successfully developed in that past that exploit the respective peptide-binding grooves of, for example, Cyclin A [344], MDM2 [224] and HSP90 [345].

We expanded the synthetic tools to produce EpCAM-like synthetic membrane protein containing the same composition as the synthetic AGR2 binding mini-protein

and the addition of transmembrane helix of EpCAM Figure 6.13). The expression of the synthetic membrane protein containing wild-type AGR2 peptide motifs (synMEM^{wt}) follows the same rule as the synPRO^{wt}, where the expression is dependent on AGR2 expression (Figure 6.14). Remarkably, immunofluorescence analysis showed that the synMEM^{wt} partially localised in the plasma membrane in AGR2-expressing cells MCF-7 but not AGR2-null cells (FLO-1). This data provides another strong evidence that such sequence-peptide peptide motif in binding to AGR2 that result in protein maturation and delivery to its destined location and this can be recapitulated using artificial membrane proteins. Further experiment should be done to further prove this for example by counting the number of cells with plasma membrane expression using FACS in unfixed cells. Based on the immunofluorescence studies, we only observe partial plasma membrane staining and there is a degree of puncta staining in the cytoplasm. Based on the immunofluorescence studies, we only observe partial plasma membrane staining and there is a degree of puncta staining in the cytoplasm. We thought that this could be due to the incorporation of the KDEL sequence which is a strong ER localisation signal and this sequence also is absent in membrane protein like EpCAM. We hypothesized that if we remove this canonical ER retention motif KDEL we would expect more plasma membrane distribution. It is also interesting to assess the effect of reconstituting AGR2 and synMEM^{wt} in FLO-1 cells (e.g whether AGR2 can help in maturation and trafficking of synMEM^{wt} to the plasma membrane, or perhaps help the production of (synMEM^{wt}) like we did with EpCAM in Chapter 4.

All in all, we were able to isolate high-affinity scFv fragments that bind to AGR2 which can be further manipulated towards more innovative biologic drugs of mammalian disease. We also demonstrated using a cell-based assay that the AGR2 sequence-specific peptide binding activity is to drive an interaction with proteins destined for the secretory pathway. AGR2 can function as an oncogene by virtue of binding to and regulate intracellular localization of not just EpCAM but synthetic membrane protein containing such sequence-specific peptide motif.

CHAPTER 7: Summary and Future Perspectives

AGR2 protein has emerged in the past ten years as a major pro-oncogenic and metastatic player in cancer largely due to the emergence of clinical-OMICS [94] [116]. In human, AGR2 was first discovered as an overexpressed gene in ER-positive breast cancer cell lines [126] and has since been shown to be overexpressed in breast cancers and many other adenocarcinomas, including colorectal, oesophageal, lung, ovarian, pancreatic, and prostate carcinomas [94]. The mechanism of how AGR2 involves in cancer and how it is regulated is vaguely understood. AGR2 can involve in the pathways for endoplasmic reticulum stress, protein folding, transcription regulation, and exosome formation, as well as in biologically important pathways including limb regeneration, inflammatory bowel disease, and asthma [54, 60, 117]. There are small genomic variants found, and mostly AGR2 is associated with overexpression, suggesting its expression is vital for cancer cell survival. Indeed, its expression is linked to resistance to some current chemotherapeutic compounds, predicts poor prognosis and has a suppressive effect on the p53 response to DNA damage. AGR2 also involve in tumour dissemination as it has been found to induce metastasis; found in circulating tumour cells (CTC) from breast, colon and prostate cancer patients [346]; found in higher concentration in the blood of patients with ovarian cancer [135] and ectopic endometriotic lesions [347]. Very little is known about the dominant pathways through which AGR2 signals in cancer, and therefore we are currently not able to embed AGR2 in any channel. Current objectives of this emerging AGR2 field, therefore, are to characterise and identify the protein interactions that mediate its function which include identification of the signalling pathways through which it can stimulate cell growth or drug-resistance [185, 186], and determining whether these pathways are “druggable”[187].

We aim to investigate the biology of AGR2 using a combination of interactomics to identify functionally relevant binding partners. We reasoned that expanding the AGR2 interactome and placing “AGR2 node” in the complex network of cellular interactions could help to unravel its functions in normal and tumour samples. We began by addressing an intriguing and ill-defined core function of AGR2

in which AGR can bind sequence-specifically to a peptide isolated from a massive peptide-phage library [209]. The peptide can affinity purify monomeric and dimeric AGR2 from crude lysates which suggest that the peptide is highly-specific and this feature of AGR2 can selectively drive some of its protein-protein interaction functions. Using hydrogen-deuterium exchange mass spectrometry, the peptide binding site was found to be in the exposed unstructured turn in the region of aa 131-135 on AGR2. Based on the previous multiple sequence alignment of AGR2 homologs and related families [57, 61] including our own alignment showed that this region is highly conserved, indicating the likely importance of binding to the peptide interface for protein function. We then developed a consensus site (Tx[IL][YF][YF]) for its activity, and use this knowledge to identify potential cellular proteins that harbour this motif as potential AGR2 client proteins. Interestingly, transmembrane proteins predominated containing such a motif using web-based informatics approach. The implication of this data is that the specific docking site of AGR2 is used to interact with client proteins that enter the endoplasmic reticulum; in which we have validated one key protein-protein interaction, represented by AGR2 and the oncogenic receptor EpCAM using biochemical assays and mutagenesis. We suggest that AGR2 has a novel function in which the AGR2 sequence-specific peptide binding is driving the interaction with proteins that are dominantly destined for the secretory pathway (Figure 7.1). In cancer cells overexpression of AGR2 may suggest that AGR2 is needed to promote cell survival by allowing the cell to withstand excess protein production associated with transformed cells and many of these proteins could be largely membrane proteins. In addition, many of the proteins identified by our database mining analysis have functions associated with cell growth, metastasis, and ER-stress responses. AGR2 can also be found in other cellular localisations such as in the cytoplasm and nucleus of some cell types, indicating the contribution of AGR2 to cell proliferation, survival, metastasis and drug resistance may be complex and mutli-faceted. Another member of the lab has also validated protein-protein interaction with another membrane-bound Meckelin protein (MKS3) which contains more divergent motif (TPIFY) using immunoprecipitation and colocalisation analyses which further supported the validity and specificity of the peptide motif, and this peptide is not simply due to a bias selection during peptide-phage display. Unfortunately, we could not find any

overlapping protein with the previously published AGR2 interacting partners, indicating a different AGR2-dependent processing or the identification could be missed using the classical approaches such as yeast two-hybrid and immunoprecipitation. In a yeast two-hybrid for example, proteins of the secretory compartments or integral membrane proteins often do not interact in the environment of the yeast nucleus, and it might not be possible to identify interaction that is dependent on post-translational modifications [348]. Thus, our data provide a different way of identifying interacting proteins using in which we used data-driven approach. Future cell-based assay developments should be directed at determining the impact on understanding how the EpCAM docking site (aa 247-251) facilitates its maturation or trafficking and how the AGR2-peptide binding motif impacts upon client protein maturation. For example, it would be interesting to generate a gain-of-function $AGR2^{S134A}$ (AGR2 docking site mutant that showed gain-of-function binding to EpCAM) or cysteine-mutated EpCAM cell line using CRISPR/Cas9 gene editing cancer to measure changes influx of EpCAM receptor maturation because of such mutations. Since both AGR2 and EpCAM can be secreted, it would also be interesting to check if AGR2 has an extracellular function in the future since the peptide motif in EpCAM is extracellular (also extracellular in MKS3). The recent establishment of AGR2 crystal structure showed that AGR2 exist in homodimer [61]. In addition, our group showed that the single cysteine of AGR2 is responsible for regulating oxidation state and determine its monomeric-dimeric state [99]. It remains unclear what role of AGR2 monomer-dimer equilibrium plays in the normal and pro-oncogenic properties of AGR.Thus, follow-up studies could also be directed to assess the AGR2 dimer/monomer equilibrium, and how destabilisation will affect characteristics such as interaction partner binding.

We further characterised the AGR2: EpCAM interaction using cell-based assays to understand how these two proteins can promote tumorigenesis since both proteins are major players in cancer development. It is also important to note that there is a targeted cancer immunotherapy drug targeting EpCAM (Catumaxomab, a monoclonal bispecific trifunctional antibody) which was approved to be used in the European Union [217]. Thus, studying EpCAM signalling and its maturation are

physiologically significant which can lead to the identification of converged pathway interconnecting AGR2 and EpCAM and may increase the future efficacy of drug development (e.g. EpCAM cell surface representation is important for recognition by an antibody for tumour clearance). We demonstrated that AGR2 and EpCAM could bind in cells using PLA and confocal microscopy co-localisation analysis. Mutation of the motif on EpCAM^{Y251A} (TLIYY to TLIYA) impaired its membrane localisation in MCF-7 cells (AGR2+/EpCAM +) and remained in the cytosol which implied that the AGR2 docking site is important in cells for the appropriate trafficking of EpCAM to its destination. We also presented a mechanism for controlling EpCAM-mediated signalling using reconstitution studies in FLO-1 cells (AGR2-/EpCAM-). AGR2 can dictate signalling activity by determining whether EpCAM is delivered to the plasma membrane from the secretory pathway. Also, AGR2 could elevate EpCAM production as confirmed by fluorescence microscopy and western blot analysis. Evaluation of AGR2 and EpCAM clinical tissues of oesophageal adenocarcinomas exhibited that the two proteins were highly co-expressed and showed significant positive correlation suggesting physiologically relevant biomarker candidates for this type of cancer for early diagnosis and for future therapeutic intervention. As discussed earlier, the next step is to establish whether EpCAM glycosylation has any effect in binding to pro-oncogenic AGR2 or whether AGR2 has any role in the glycosylation of EpCAM in the future. Our results thus far showed that AGR2 binds EpCAM in the ER, therefore it remains to be determined whether AGR2 can interact with EpCAM post-ER or at later steps in the secretory pathway (Figure 7.1). Intriguingly, a recent study described the unexpected gain-of-function of an extracellular AGR2 (eAGR2) as a regulator of epithelial tumorigenesis and morphogenesis [180] suggesting another role of AGR2 in cancer development. Since EpCAM can also be cleaved to produce extracellular fragment (EpEx) and in this fragment that EpCAM harbours AGR2 peptide motif, it would be interesting to study the relationship of the extracellular role of AGR2 and EpCAM in tumour signalling.

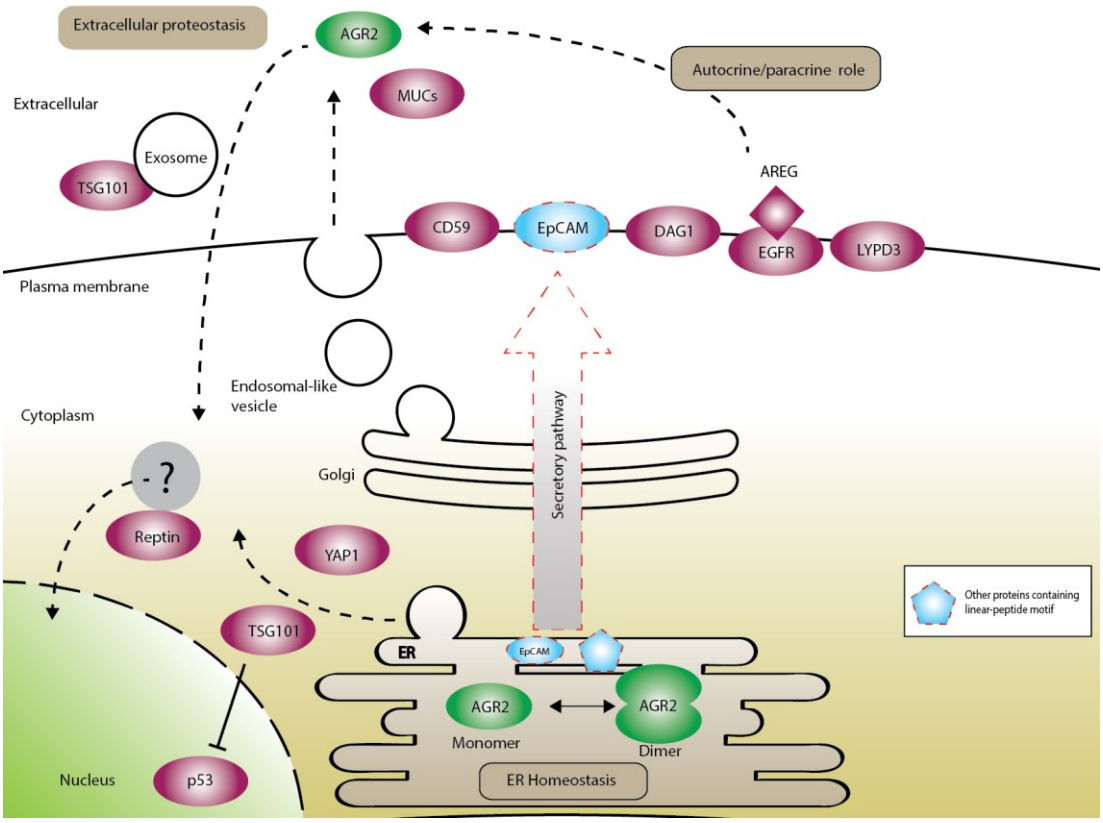


Figure 7.1 Model of AGR2 role in the secretory landscape.
AGR2 has a novel function in which AGR2 binds to a subset of client proteins containing AGR2 linear-peptide motif that are destined for secretory pathway. Many proteins that harbour the linear-peptide motif belong to membrane type proteins. In this thesis, we have validated one member of this type of protein represented by oncogenic EpCAM. Our data showed that AGR2: EpCAM interaction occurs in the ER, the first compartment of the secretory pathway, possibly through its PDI family chaperone function in which AGR2 help in folding client protein like EpCAM. It remains to be determined if the interaction occurs at later steps in the secretory pathway. Our database mining approach expands AGR2 interactome and suggests a biological role for AGR2 in the secretory pathway.

During this study, we attempted to make a mouse monoclonal antibodies targeting EpCAM. We sent the purified codon-optimised extracellular part of EpCAM (His-EpEx) for immunisation in mice. Three initial clones showed a weak affinity binding to EpCAM in western blot, and these were selected and grown further. Eventually, only one clone can be produced, however, this clone of EpCAM monoclonal antibody was not able to detect purified EpCAM nor EpCAM in cell lysates in western blot analysis (data not shown). Currently, we are in the process of characterising the localisation of AGR2-CHERRY and EpCAM-GFP using live-cell fluorescence imaging. We briefly showed in Chapter 5, that AGR2 movement resembles secretory vesicle and has a high rate of mobility throughout the cells as analysed with live-cell fluorescence imaging, which quite unusual for ER-resident protein. Our preliminary data exhibited that, when AGR2-CHERRY and EpCAM-GFP were co-expressed in cells, the movement of AGR2-CHERRY was reduced suggesting that EpCAM impacted on AGR2 signalling and there is a positive feedback interaction exists between the two pathways. This also further supported that AGR2 and EpCAM interact in cells. However, we are yet to find a mechanism for such observation.

We sought to generate isogenic cell lines that carry AGR2 knockout using the latest CRISPR/Cas9 technology. This study aimed to use this isogenic oesophageal cell model (AGR2 knockout vs. wt-AGR2) to screen for global proteomic changes using quantitative proteomics which serve as another paradigm to identify potential AGR2 interacting partners and to define AGR2 regulatory network in cancer signalling. However, we were unable to generate a cell line with AGR2 knockout after two different attempts. We postulated that AGR2 edited knockout cell line was lethal which was supported by a few pieces of evidence in mouse knockout models showing lethality as a result of AGR2 knockout as discussed previously [60, 63, 289]. We then opted to reconstitute AGR2-null cells, FLO-1 using transient AGR2 overexpression and process for mass-spectrometry analysis. TMT mass spectrometry quantitation showed that the proteome of reconstituted wt-AGR2 cells exhibited many downregulated proteins while the AGR2-KDEL showed many upregulated proteins suggesting that the two AGR2 variants have specific distinct signalling pathways.

There were overlapping proteins between the two AGR2 variants expressing cells which can be potential AGR2 interacting partners and can be integrated into AGR2 'signalling pathways'. Both AGR2 reconstituted cells showed enrichment in exosomal and lysosomal proteins suggesting AGR2 role in cancer signalling in this oesophageal cancer model and this also complemented our current hypothesis that an ER-resident protein AGR2 involve in the secretion of proteins destined for secretory pathway. We chose one protein, DCD, which is a secreted protein and has been implicated in cancer, as a potential AGR2 interacting protein. Our initial data showed that the level of DCD was downregulated in both AGR2-expressing cells compared to non-reconstituted cells, which contradicted the mass spectrometry data. The fact that DCD is rarely found in its full-length form due to high-susceptibility cleavage by proteases to form functional peptides, and the lack of antibodies, complicates our analysis to detect DCD changes in our cell models. DCD was also found in another proteomics screen in our lab SWATH mass spectrometry analysis of ELMO1 pull-down from FLO-1 and OE19 cells (Ali Alsaadi thesis 2016, and personal communication). ELMO1 was shown to form protein-protein interaction complex with AGR2 as well as DCD in cells using PLA. Since we do not have evidence that AGR2 interacts with DCD, we speculate that AGR2 may bind DCD through the interaction with ELMO1 or with one of ELMO1 signalling proteins. The next steps of this study are to define the mechanism whereby AGR2, predominantly an ER-resident molecular chaperone, promotes the expression and secretion of DCD and then to link to changes in key cellular phenotypes (e.g., does AGR2 affect DCD proteolysis proteolytic processing). We recently obtained some DCD drugs in which we can test on our cell models. For example, we asked whether AGR2 expression in FLO-1 cells (since AGR2 increase production of DCD) can sensitise DCD to these drugs. We also aimed to repeat the experiment by performing protein glycosylation-targeting enrichment prior to mass spectrometric analysis to test our hypothesis that AGR2 can bind and mature membrane-bound proteins. In addition, most of the membrane or transmembrane proteins identified from database mining in Chapter 3 are known to be post-translationally modified particularly glycosylation and most contained a large number of cysteine residues.

Along with the identification of AGR2 interacting proteins, synthetic tools were devised and engineered to dissect the pro-oncogenic AGR2 function further. This study, using synthetic scFv phage library, we have successfully isolated 5 different classes of scFv fragments that bind to both full-length AGR2 and AGR2 peptides. One particular scFv fragment (scFv4) bind with high-affinity to the N-terminal of AGR2. Expression of the scFv4 in mammalian cells showed that its expression are regulated by AGR2, but AGR2 expression itself remain unchanged. Therefore, next study can be focused on finding the mechanism of how this AGR2-specific scFv can affect AGR2 expression. Currently, we are in the process of purifying the AGR2-binding scFv4 from bacteria so that we can further characterise the antibody (e.g., how scFv binding to AGR2 influence AGR2 structure, and affect AGR2 interactome) and determine its pharmacokinetics properties. Experiments are underway to clone scFv into an imaging vector that fused the scFv with EGFP protein, so we can analyse its localisation, and check if the expression has an impact on AGR2 expression or AGR2 interacting protein in cells. This scFv can be potentially translated for the specific probe for *in-vivo* fluorescence imaging, for example, use in the detection of cancer-specific cancer since AGR2 was selected as biomarkers for some cancer types. We also developed a synthetic mini protein containing AGR2-peptide docking motif (synPRO^{wt}) to show the specificity of the peptide in cells. The mini-protein expression in mammalian cells was negatively regulated by AGR2 and its expression stabilised AGR2 protein levels. Also, PLA showed that the AGR2 and synPRO^{wt} form protein-protein interaction in cells. These findings further supported our *in-vitro* data that AGR2 has a sequence-specific peptide binding activity that drives (some of) its protein-protein interaction. Finally, we engineered synthetic membrane protein (synMEM^{wt}) containing AGR2 mini-protein and EpCAM transmembrane domain. Expression of the synthetic membrane protein was also regulated by AGR2 expression. Importantly, in the presence of endogenous AGR2, the expression of the synMEM^{wt} was seen to localise partially in the plasma membrane as analysed by immunofluorescence, suggesting that AGR2 recognise the sequence specific peptide which then channels the protein to secretory pathway and to its final destination. Future cell-based assays can be performed to check if synMEM^{mut} plasma membrane localisation can be rescued by ectopically expressing AGR2 in AGR2-null cells (FLO-

1). This will further support our data that there is a sequence-peptide peptide docking motif in AGR2 activity. The specific docking site of AGR2 is used to interact with client proteins that enter the endoplasmic reticulum which is then processed for proper maturation and delivery to its destined location.

In conclusion, our data highlight an unusual feature of a "molecular chaperone" class of proteins that exist in the endoplasmic reticulum (sequence-specific peptide binding). We discovered that proteins containing the consensus AGR2-peptide docking motifs were highly enriched in transmembrane proteins. We suggest that AGR2 has a novel intracellular function in which the AGR2 sequence-specific peptide binding is driving the interaction with a subset of client proteins while they are trafficked into the secretory pathway. We also provided a working model of how AGR2 can function as an oncogene by virtue of binding to and regulating intracellular localisation of one member of membrane proteins which is an oncogenic receptor EpCAM.Such sequence-specific binding can also be recapitulated using synthetic membrane protein containing such a motif. We presented data from proteomics screens suggesting that AGR2 showed upregulation for proteins involved in secretory pathways. We also developed synthetic tools such as high-affinity scFv fragment targeting AGR2 and mini-protein containing AGR2-peptide docking motifs that can be further manipulated to target AGR2 signalling pathway.

CHAPTER 8: Appendices

8.1. Appendix 1

Raw graphical data summarising AGR2 ligand binding (peptide A4) effects on AGR2 conformation analysed by hydrogen-deuterium exchange mass spectrometry.

8.2. Appendix 2

Protein hits list from database mining using AGR2 consensus peptide binding motif Tx[IL][YF][YF] in ScanProsite (http://prosite.expasy.org/scanprosite/). Hits were matched with UniprotKB entries and the motif hits were listed in bold fonts.

sp|P48023|TNFL6_HUMAN **TgLYF**
sp|P32248|CCR7_HUMAN **TyIYF**
sp|Q9Y2K9|STB5L_HUMAN **TaLYF**
sp|P30874|SSR2_HUMAN **TfIYF**
sp|P59536|T2R41_HUMAN **TlLFF**
sp|O76099|OR7C1_HUMAN **TgIFF**
sp|Q7Z402|TMC7_HUMAN **TsLFY**
sp|O43314|VIP2_HUMAN **TrLYF**
sp|O75691|UTP20_HUMAN **TdLYY**
sp|Q00975|CAC1B_HUMAN **TtLYF**
sp|Q8NGK2|O52B4_HUMAN **TqLFF**
sp|Q5VZL5|ZMYM4_HUMAN **TvLFY**
sp|Q5VZL5|ZMYM4_HUMAN_2 **TlLFF**
sp|P51582|P2RY4_HUMAN **TlIYY**
sp|Q7Z407|CSMD3_HUMAN **TsLYF**
sp|Q7L1T6|NB5R4_HUMAN **TgLYY**
sp|Q15620|OR8B8_HUMAN **TqLFF**
sp|Q92581|SL9A6_HUMAN **TgLFF**
sp|P01774|HV313_HUMAN **TaLYY**
sp|Q9Y4I1|MYO5A_HUMAN **TkIFF**
sp|Q92806|KCNJ9_HUMAN **TwLFF**
sp|Q9P0J1|PDP1_HUMAN **TqLFF**
sp|Q9H015|S22A4_HUMAN **TsLFF**
sp|Q13790|APOF_HUMAN **TaLYY**
sp|Q9BRB3|PIGQ_HUMAN **TaLYY**
sp|Q96RW7|HMCN1_HUMAN **TeLFY**
sp|Q9P2E3|ZNFX1_HUMAN **TqIFF**
sp|Q5HYA8|MKS3_HUMAN **TpIFY**
sp|Q5HYA8|MKS3_HUMAN_2 **TvLFF**
sp|Q8TDX9|PK1L1_HUMAN **ThLFF**
sp|Q8IX03|KIBRA_HUMAN **TdLYY**
sp|A6NF89|OR6C6_HUMAN **TqLFF**
sp|Q8NGN0|OR4D5_HUMAN **TqLFF**
sp|Q5BJE1|CC178_HUMAN **TmIFY**
sp|Q13948|CASP_HUMAN **TaLFY**
sp|Q9BXB1|LGR4_HUMAN **TlIFF**
sp|Q9NP91|S6A20_HUMAN **TqIFF**
sp|P47881|OR3A1_HUMAN **TqLFF**
sp|Q8IZU8|DSEL_HUMAN **TnLFY**
sp|Q32P28|P3H1_HUMAN **TpLYF**
sp|P51681|CCR5_HUMAN **TgLYF**
sp|Q96R28|OR2M2_HUMAN **TfLFF**
sp|Q96R28|OR2M2_HUMAN_2 **TqIFF**
sp|Q86W33|TPRA1_HUMAN **TfLYF**
sp|Q9Y2G2|CARD8_HUMAN **TlIYY**
sp|Q9BVV8|CS024_HUMAN **TaLYF**
sp|Q5T5X7|BEND3_HUMAN **TeIYF**
sp|Q8WTR4|GDPD5_HUMAN **TwLYF**
sp|Q6QNK2|AGRD1_HUMAN **TlLYY**
sp|P01779|HV318_HUMAN **TaLYY**
sp|Q86XM0|CTSRD_HUMAN **TyIYY**
sp|Q9H330|TM245_HUMAN **TiLFY**
sp|Q9H330|TM245_HUMAN_2 **TtLFY**
sp|P11717|MPRI_HUMAN **TlIYF**
sp|P11717|MPRI_HUMAN_2 **TaIFF**
sp|Q8NG92|O13H1_HUMAN **TpIYF**
sp|P48051|KCNJ6_HUMAN **TwLFF**
sp|Q5JTV8|TOIP1_HUMAN **TlIFY**
sp|Q96DT5|DYH11_HUMAN **TpIFF**
sp|Q8NBJ9|SIDT2_HUMAN **TqLYY**
sp|Q8IWT6|LRC8A_HUMAN **TqLFY**
sp|Q96RV3|PCX1_HUMAN **TvLFF**
sp|Q8N394|TMTC2_HUMAN **ThIFY**
sp|Q8N394|TMTC2_HUMAN_2 **TnLFF**
sp|Q12767|TMM94_HUMAN **TvLFF**
sp|O95461|LARGE_HUMAN **ThLYF**
sp|Q15035|TRAM2_HUMAN **TiLFY**
sp|P47884|OR1D4_HUMAN **TdLFF**
sp|P47884|OR1D4_HUMAN_2 **TqLYF**
sp|Q96A58|RERG_HUMAN **TeIFY**
sp|Q7RTT9|S29A4_HUMAN **TlIFF**
sp|Q86XW9|TXND6_HUMAN **TfLFY**
sp|Q9UBU6|FA8A1_HUMAN **TlLFF**
sp|P42898|MTHR_HUMAN **TqLFF**
sp|A5D6W6|FITM1_HUMAN **TvIYF**
sp|Q12864|CAD17_HUMAN **TgLFF**
sp|Q7Z7L7|ZER1_HUMAN **TnIFY**
sp|P16422|EPCAM_HUMAN **TlIYY**
sp|Q6NVV3|NIPA3_HUMAN **TpIYY**
sp|Q96GW9|SYMM_HUMAN **TpIFY**

sp|Q9UN76|S6A14_HUMAN **TqIFY**
sp|O76027|ANXA9_HUMAN **TpLYF**
sp|O95800|GPR75_HUMAN **TlIFF**
sp|P01009|A1AT_HUMAN **TnIFF**
sp|P01009|A1AT_HUMAN_2 **TaIFF**
sp|Q8NHV5|CP052_HUMAN **TtLFF**
sp|Q86X27|RGPS2_HUMAN **TqLFY**
sp|Q9ULZ2|STAP1_HUMAN **TtLFF**
sp|Q9H9C1|SPE39_HUMAN **TtLFY**
sp|Q9Y5Z7|HCFC2_HUMAN **TrLYF**
sp|Q9BRQ3|NUD22_HUMAN **TgIFF**
sp|Q9UKZ4|TEN1_HUMAN **TsLYY**
sp|Q8N6P7|I22R1_HUMAN **TeLYY**
sp|Q9NZM6|PK2L2_HUMAN **TfIFF**
sp|Q9HB14|KCNKD_HUMAN **TiLFF**
sp|Q8IVQ6|ZDH21_HUMAN **TgLFY**
sp|O00468|AGRIN_HUMAN **TrIFF**
sp|P51817|PRKX_HUMAN **TgLFY**
sp|Q99504|EYA3_HUMAN **ThLFF**
sp|Q9Y2P8|RCL1_HUMAN **TtLYY**
sp|Q9NQX4|MYO5C_HUMAN **TkIFF**
sp|O95185|UNC5C_HUMAN **TqIYF**
sp|Q9BZJ4|S2539_HUMAN **TaIYF**
sp|O00462|MANBA_HUMAN **TiLFY**
sp|Q8NGX8|OR6Y1_HUMAN **TqLYF**
sp|Q96KS0|EGLN2_HUMAN **TcIYY**
sp|P10600|TGFB3_HUMAN **TiLYY**
sp|Q8IWT3|CUL9_HUMAN **TcLFY**
sp|Q12809|KCNH2_HUMAN **TaLYF**
sp|Q12809|KCNH2_HUMAN_2 **TaLYF**
sp|P58182|O12D2_HUMAN **TyLFF**
sp|A7MBM2|DISP2_HUMAN **TlIFF**
sp|A7MBM2|DISP2_HUMAN_2 **TqLFY**
sp|Q6ZTZ1|MSD1_HUMAN **TpLYF**
sp|Q96JM2|ZN462_HUMAN **TeLYY**
sp|Q9UKT6|FXL21_HUMAN **ThLYF**
sp|Q8WUX2|CHAC2_HUMAN **TvIFY**
sp|Q9HBW0|LPAR2_HUMAN **TrIFF**
sp|Q15061|WDR43_HUMAN **TgLYF**
sp|O00167|EYA2_HUMAN **ThLFF**
sp|P10747|CD28_HUMAN **TdIYF**
sp|Q8TAD4|ZNT5_HUMAN **TvIFF**
sp|Q9NUE0|ZDH18_HUMAN **TgLFF**
sp|O14966|RAB7L_HUMAN **TrLYY**
sp|Q8IVW8|SPNS2_HUMAN **TaLFF**
sp|P01781|HV320_HUMAN **TaLYY**
sp|Q969K4|ABTB1_HUMAN **TpLYY**
sp|Q92636|FAN_HUMAN **TnLYF**
sp|P24928|RPB1_HUMAN **TrLFY**
sp|P24928|RPB1_HUMAN_2 **TaIYY**
sp|Q6PKC3|TXD11_HUMAN **TvLFF**
sp|Q7L211|ABHDD_HUMAN **TiIYF**
sp|P78527|PRKDC_HUMAN **TeLYF**
sp|Q9Y345|SC6A5_HUMAN **TqIFF**
sp|O76003|GLRX3_HUMAN **TfLFF**
sp|Q8IZ69|TRM2A_HUMAN **TcLYF**
sp|Q8WXJ9|ASB17_HUMAN **TpLFY**
sp|Q5T9L3|WLS_HUMAN **TvIFF**
sp|Q8NFP9|NBEA_HUMAN **TeIYF**
sp|Q5TZK3|FAM74_HUMAN **TwLFF**
sp|Q9NTJ3|SMC4_HUMAN **TpLYF**
sp|O43149|ZZEF1_HUMAN **TeLFF**
sp|Q9NSD5|S6A13_HUMAN **TqIFF**
sp|Q92484|ASM3A_HUMAN **TnLYY**
sp|Q96PE1|AGRA2_HUMAN **TwIYF**
sp|Q03721|KCNC4_HUMAN **TmIYY**
sp|Q5GH70|XKR9_HUMAN **TyLFY**
sp|Q8N9W4|GG6L2_HUMAN **TaLYY**
sp|Q9Y320|TMX2_HUMAN **TiLFF**
sp|Q7Z3Z3|PIWL3_HUMAN **TrLFY**
sp|Q8WXK1|ASB15_HUMAN **TaLYF**
sp|Q9ULG1|INO80_HUMAN **TvIFY**
sp|P48065|S6A12_HUMAN **TqIFF**
sp|Q14515|SPRL1_HUMAN **TgLFF**
sp|Q9NT68|TEN2_HUMAN **TsLYY**
sp|Q13459|MYO9B_HUMAN **TgLFY**
sp|Q9H4Q4|PRD12_HUMAN **TsIFY**
sp|Q9Y2P4|S27A6_HUMAN **TnLFY**
sp|Q6ZN44|UNC5A_HUMAN **TqIFF**
sp|Q96RD0|OR8B2_HUMAN **TrLFF**
sp|P16473|TSHR_HUMAN **TdIFF**
sp|Q8IYP9|ZDH23_HUMAN **TaLFY**
sp|Q8N3K9|CMYA5_HUMAN **TgIYF**
sp|P31641|SC6A6_HUMAN **TqIFF**
sp|Q9UJ42|GP160_HUMAN **TiLYF**
sp|Q8TBP6|S2540_HUMAN **TvIYF**
sp|Q99805|TM9S2_HUMAN **TiLYF**
sp|P15309|PPAP_HUMAN **TeLYF**
sp|Q9P1W3|CSC1_HUMAN **TfLFF**
sp|Q9P1W3|CSC1_HUMAN_2 **TrLFF**
sp|Q00722|PLCB2_HUMAN **TdIFF**
sp|Q86U70|LDB1_HUMAN **TeLYY**
sp|Q8TDM6|DLG5_HUMAN **TdIFY**
sp|Q9NX61|T161A_HUMAN **TrLYF**
sp|Q8WWT9|S13A3_HUMAN **TiLFF**
sp|Q6ZMK1|CYHR1_HUMAN **TrLYY**
sp|Q7KZN9|COX15_HUMAN **TvLYF**
sp|Q8NGR6|OR1B1_HUMAN **TlLFF**
sp|Q8IVB4|SL9A9_HUMAN **TgLFF**
sp|Q96Q35|AL2SB_HUMAN **TnLYY**
sp|O75503|CLN5_HUMAN **TgIYY**
sp|Q15617|OR8G1_HUMAN **TqLYF**
sp|Q8NGW6|OR6K6_HUMAN **TpLYF**
sp|P30531|SC6A1_HUMAN **TqIFF**
sp|Q8WXI7|MUC16_HUMAN **TsIFF**
sp|Q3ZAQ7|VMA21_HUMAN **TlLFF**
sp|O60412|OR7C2_HUMAN **TqIFF**
sp|Q14202|ZMYM3_HUMAN **TdLYY**
sp|Q8IWU2|LMTK2_HUMAN **TsIFY**
sp|Q8NGD1|OR4N2_HUMAN **TqLFF**
sp|Q7L0X2|ERIP6_HUMAN **TqIFY**
sp|Q5JRS4|O10J3_HUMAN **TqLFF**
sp|Q9ULV0|MYO5B_HUMAN **TkIFF**
sp|Q7Z3B1|NEGR1_HUMAN **TsIFY**
sp|Q14494|NF2L1_HUMAN **TgLFF**
sp|Q5VWN6|F208B_HUMAN **TfLFY**

sp|P80303|NUCB2_HUMAN **TgLYY**
sp|P19256|LFA3_HUMAN **TsIYF**
sp|Q96GJ1|TRM2_HUMAN **TsLYF**
sp|Q14289|FAK2_HUMAN **TlLYF**
sp|P51168|SCNNB_HUMAN **TsIFY**
sp|Q9BZZ2|SN_HUMAN **TsLFF**
sp|Q96FC7|PHIPL_HUMAN **TnLYF**
sp|Q5XXA6|ANO1_HUMAN **TpIFY**
sp|P08637|FCG3A_HUMAN **TgLYF**
sp|Q15619|OR1C1_HUMAN **TqLFF**
sp|Q13769|THOC5_HUMAN **TnLYY**
sp|O75015|FCG3B_HUMAN **TgLYF**
sp|Q6ZVE7|GOT1A_HUMAN **TlLYF**
sp|P35125|UBP6_HUMAN **TeLYF**
sp|Q9C0B5|ZDHC5_HUMAN **TtLFF**
sp|P12259|FA5_HUMAN **ThIYY**
sp|Q08ET2|SIG14_HUMAN **TwIYY**
sp|Q02818|NUCB1_HUMAN **TgLYY**
sp|O95754|SEM4F_HUMAN **TpIFY**
sp|Q8NGX3|O10T2_HUMAN **TqLFF**
sp|P16435|NCPR_HUMAN **TlLYY**
sp|Q8NGX9|OR6P1_HUMAN **TqLYF**
sp|P06326|HV107_HUMAN **TaIYY**
sp|Q9Y4D8|HECD4_HUMAN **TdLFY**
sp|P59551|T2R60_HUMAN **TiLFF**
sp|Q92523|CPT1B_HUMAN **TgIFF**
sp|Q9GZM6|OR8D2_HUMAN **TqLYF**
sp|Q7Z2Y8|GVIN1_HUMAN **TkLYF**
sp|Q9NX02|NALP2_HUMAN **TaLFY**
sp|Q9NXL6|SIDT1_HUMAN **TqIYY**
sp|Q9NXL6|SIDT1_HUMAN_2 **TaLFF**
sp|Q8NFQ8|TOIP2_HUMAN **TlIFY**
sp|Q9ULD8|KCNH3_HUMAN **TsLYF**
sp|Q86TU7|SETD3_HUMAN **TpLYF**
sp|P48544|KCNJ5_HUMAN **TwLFF**
sp|P23975|SC6A2_HUMAN **TqIFF**
sp|Q96CN9|GCC1_HUMAN **TaLYY**
sp|O95789|ZMYM6_HUMAN **TvLFY**
sp|P54922|ADPRH_HUMAN **TqLYY**
sp|Q8NGD2|OR4K2_HUMAN **TqIFF**
sp|Q9H6D3|XKR8_HUMAN **TiLYF**
sp|Q8IVF5|TIAM2_HUMAN **TlLFY**
sp|Q49SQ1|GPR33_HUMAN **TlLFF**
sp|Q8NGK1|O51G1_HUMAN **TqLFF**
sp|Q8IYB7|DI3L2_HUMAN **TsLFF**
sp|Q8IXE1|OR4N5_HUMAN **TqLFF**
sp|Q86WJ1|CHD1L_HUMAN **TyIYY**
sp|Q8IUA7|ABCA9_HUMAN **TfIYY**
sp|O95222|OR6A2_HUMAN **TqLYF**
sp|Q8N6D5|ANR29_HUMAN **TaLFF**
sp|Q8NG78|OR8G5_HUMAN **TqLYF**
sp|Q7Z602|GP141_HUMAN **TfLFY**
sp|Q9UHD2|TBK1_HUMAN **TeIYY**
sp|Q2KHT3|CL16A_HUMAN **TsLYY**
sp|Q9GZT9|EGLN1_HUMAN **TcIYY**
sp|P98175|RBM10_HUMAN **TgLYY**
sp|P04040|CATA_HUMAN **TpIFF**
sp|Q86XT9|TM219_HUMAN **TcLYF**
sp|Q6IEE8|SN12L_HUMAN **TkIFY**
sp|Q6AWC2|WWC2_HUMAN **TeLYY**
sp|Q5JS13|RGPS1_HUMAN **TlLYY**
sp|O43679|LDB2_HUMAN **TdLYY**
sp|Q8NFM4|ADCY4_HUMAN **TsLFF**
sp|Q9NYI0|PSD3_HUMAN **TsLYY**
sp|Q8N8Q9|NIPA2_HUMAN **TpIYY**
sp|Q96RA2|OR7D2_HUMAN **TlIYF**
sp|Q9NS40|KCNH7_HUMAN **TaLYF**
sp|Q9NS40|KCNH7_HUMAN_2 **TaLYF**
sp|Q14003|KCNC3_HUMAN **TmIYY**
sp|P48147|PPCE_HUMAN **TeIFY**
sp|Q8NGB8|O4F15_HUMAN **TqIFF**
sp|Q9H2S6|TNMD_HUMAN **TgIYF**
sp|Q9Y2L1|RRP44_HUMAN **TqLFF**
sp|P22223|CADH3_HUMAN **TkIFY**
sp|P25101|EDNRA_HUMAN **TaIFY**
sp|Q8TD84|DSCL1_HUMAN **TsLYF**
sp|P15169|CBPN_HUMAN **TyIYY**
sp|Q96QP1|ALPK1_HUMAN **TqIFY**
sp|Q9BYI3|HYCCI_HUMAN **TgIYF**
sp|Q9NS93|TM7S3_HUMAN **TeLFF**
sp|Q9P273|TEN3_HUMAN **TsLYY**
sp|Q2KJY2|KI26B_HUMAN **TaLFY**
sp|O00254|PAR3_HUMAN **TvIFY**
sp|Q8NI27|THOC2_HUMAN **TkLFY**
sp|Q8NH69|OR5W2_HUMAN **TsLFY**
sp|Q92610|ZN592_HUMAN **TlLFF**
sp|Q9UKT7|FBXL3_HUMAN **ThLYF**
sp|Q969G6|RIFK_HUMAN **TgIYY**
sp|Q9P2R6|RERE_HUMAN **TdLYF**
sp|Q8NCS7|CTL5_HUMAN **TiLFY**
sp|Q8NFA0|UBP32_HUMAN **TeLYF**
sp|O60331|PI51C_HUMAN **TdIYF**
sp|P59538|T2R31_HUMAN **TvIFF**
sp|Q9NSU2|TREX1_HUMAN **TlIFF**
sp|Q8NGB2|OR4C5_HUMAN **TlIYF**
sp|P98168|ZXDA_HUMAN **TsLFF**
sp|Q7RTY7|OVCH1_HUMAN **TvIYF**
sp|P02790|HEMO_HUMAN **TmLFF**
sp|P58170|OR1D5_HUMAN **TdLFF**
sp|P58170|OR1D5_HUMAN_2 **TqLYF**
sp|Q8NG04|S2610_HUMAN **TpLYF**
sp|Q15326|ZMY11_HUMAN **TvIFY**
sp|P0DN82|O12D1_HUMAN **ThLFF**
sp|Q6P4A7|SFXN4_HUMAN **TeIFY**
sp|P48547|KCNC1_HUMAN **TmIYY**
sp|Q96N87|S6A18_HUMAN **TqIFF**
sp|Q9UK97|FBX9_HUMAN **TpLFF**
sp|P48651|PTSS1_HUMAN **TeLFF**
sp|A6NKB5|PCX2_HUMAN **TvIFF**
sp|Q3ZCN5|OTOGL_HUMAN **TsLFF**
sp|Q99502|EYA1_HUMAN **ThLFF**
sp|Q8IZJ1|UNC5B_HUMAN **TqIYF**
sp|P19634|SL9A1_HUMAN **TvIFF**
sp|P34897|GLYM_HUMAN **TsIFF**
sp|Q15198|PGFRL_HUMAN **TyIFF**
sp|P34982|OR1D2_HUMAN **TdLFF**

sp|P34982|OR1D2_HUMAN_2 **TqLYF**
sp|O43934|MFS11_HUMAN **TvLFF**
sp|Q96SA4|SERC2_HUMAN **TlLFY**
sp|O75752|B3GL1_HUMAN **TnLFF**
sp|Q92540|SMG7_HUMAN **TtIFY**
sp|Q8NDW8|TT21A_HUMAN **TaLYY**
sp|Q96PR1|KCNC2_HUMAN **TmIYY**
sp|Q63HN8|RN213_HUMAN **TiLFF**
sp|Q14999|CUL7_HUMAN **TrLFY**
sp|Q8NEC5|CTSR1_HUMAN **TfIFF**
sp|P17017|ZNF14_HUMAN **TfIYY**
sp|Q6MZM0|HPHL1_HUMAN **TyLYF**
sp|P20042|IF2B_HUMAN **TrLYF**
sp|Q9H252|KCNH6_HUMAN **TaLYF**
sp|P25705|ATPA_HUMAN **TeLFY**
sp|Q2VPB7|AP5B1_HUMAN **ThLFY**
sp|O95243|MBD4_HUMAN **TsLYF**
sp|O95677|EYA4_HUMAN **ThLFF**
sp|Q92485|ASM3B_HUMAN **TnLYY**
sp|Q8NGC2|OR4E2_HUMAN **TqLFF**
sp|Q08334|I10R2_HUMAN **TlLFF**
sp|Q8NDA8|MROH1_HUMAN **TcLFY**
sp|P53370|NUDT6_HUMAN **TgLFY**
sp|P48029|SC6A8_HUMAN **TqIFF**
sp|Q9UFH2|DYH17_HUMAN **TsIFF**
sp|Q9UFH2|DYH17_HUMAN_2 **TaLFY**
sp|P16410|CTLA4_HUMAN **TlLFF**
sp|Q2TAZ0|ATG2A_HUMAN **TdLFF**
sp|Q7Z589|EMSY_HUMAN **TgLFY**
sp|Q04671|P_HUMAN **TlLFF**
sp|Q96KA5|CLP1L_HUMAN **TnLYF**
sp|O43396|TXNL1_HUMAN **TfLFF**
sp|Q8WX94|NALP7_HUMAN **TaLFY**
sp|Q96AX1|VP33A_HUMAN **TlIFF**
sp|Q96RC9|OR8B4_HUMAN **TqLFF**
sp|O43930|PRKY_HUMAN **TgLFY**
sp|O60337|MARH6_HUMAN **TpLFY**
sp|P35503|UD13_HUMAN **TqLYF**
sp|Q66K64|DCA15_HUMAN **TkLYY**
sp|P34896|GLYC_HUMAN **TsIFF**
sp|Q9HB55|CP343_HUMAN **TiLFY**
sp|A0FGR9|ESYT3_HUMAN **TkLYF**
sp|Q16134|ETFD_HUMAN **TgIFY**
sp|Q96QB1|RHG07_HUMAN **TlLYF**
sp|Q9Y6V0|PCLO_HUMAN **TsLYF**
sp|Q01118|SCN7A_HUMAN **TfIFF**
sp|Q6IF36|OR8G2_HUMAN **TqLYF**
sp|P08F94|PKHD1_HUMAN **TgLFY**
sp|P78509|RELN_HUMAN **TaLYF**
sp|Q86WS3|OOSP2_HUMAN **TeLYF**
sp|Q96FC9|DDX11_HUMAN **TkIYY**
sp|Q92502|STAR8_HUMAN **TlLYF**
sp|Q15848|ADIPO_HUMAN **TkIFY**
sp|Q9NXG6|P4HTM_HUMAN **TvLFY**
sp|P20807|CAN3_HUMAN **TsLFY**
sp|Q9H244|P2Y12_HUMAN **TvLFF**
sp|Q9H6K1|CF106_HUMAN **TgLYY**
sp|Q8N434|SVOPL_HUMAN **TaLFF**
sp|Q96P63|SPB12_HUMAN **TiLFY**
sp|Q9NVH6|TMLH_HUMAN **TtLFF**
sp|Q86WA9|S2611_HUMAN **TsLFY**
sp|Q6N022|TEN4_HUMAN **TsLYY**
sp|Q8NGF8|OR4B1_HUMAN **TqIFF**
sp|Q6PFW1|VIP1_HUMAN **TrLYF**
sp|Q9BYC5|FUT8_HUMAN **TlLFY**
sp|Q9BYC5|FUT8_HUMAN_2 **TdLYY**
sp|Q99490|AGAP2_HUMAN **TaLFY**
sp|O60359|CCG3_HUMAN **TlLFF**
sp|P48595|SPB10_HUMAN **TiLFY**
sp|Q9NS28|RGS18_HUMAN **TlLFF**
sp|Q8IZF0|NALCN_HUMAN **TlIFF**
sp|Q7RTV0|PHF5A_HUMAN **TdLFY**
sp|Q8IYM2|SLN12_HUMAN **TkIFY**
sp|Q8TDJ6|DMXL2_HUMAN **TeIFF**
sp|Q8TDJ6|DMXL2_HUMAN_2 **TrLYF**
sp|P08842|STS_HUMAN **TlIYF**
sp|Q9Y6J0|CABIN_HUMAN **TcLYF**
sp|Q92561|PHYIP_HUMAN **TnLYF**
sp|P00813|ADA_HUMAN **TiLYY**
sp|P52756|RBM5_HUMAN **TgLYY**
sp|P31645|SC6A4_HUMAN **TvIYF**
sp|O75175|CNOT3_HUMAN **TyIYF**
sp|P52630|STAT2_HUMAN **TmLFF**
sp|Q8NG81|OR2M7_HUMAN **TfLFF**
sp|Q8NG81|OR2M7_HUMAN_2 **TqIFF**
sp|Q8N302|AGGF1_HUMAN **TgLYF**
sp|Q8N302|AGGF1_HUMAN_2 **TgIYY**
sp|Q96R27|OR2M4_HUMAN **TqIFF**
sp|Q9NRJ2|GSAS1_HUMAN **TlLFY**
sp|O75923|DYSF_HUMAN **TlIFY**
sp|Q15743|OGR1_HUMAN **TsIYF**
sp|Q8N371|KDM8_HUMAN **TeLFY**
sp|Q8NGG6|OR8BC_HUMAN **TqLFF**
sp|Q96T83|SL9A7_HUMAN **TaLFF**
sp|Q9H6Z9|EGLN3_HUMAN **TcIYY**
sp|Q8IU68|TMC8_HUMAN **TyLFY**
sp|P15559|NQO1_HUMAN **TpLYF**
sp|Q9HB15|KCNKC_HUMAN **TiLFF**
sp|P98169|ZXDB_HUMAN **TsLFF**
sp|Q8NG83|OR2M3_HUMAN **TfLFF**
sp|Q8NG83|OR2M3_HUMAN_2 **TqIFF**
sp|O75445|USH2A_HUMAN **TgIFY**
sp|P61812|TGFB2_HUMAN **TiLYY**
sp|Q9H2Y9|SO5A1_HUMAN **TpIYF**
sp|Q8NGW1|OR6B3_HUMAN **TqLYF**
sp|O94966|UBP19_HUMAN **TrLFF**
sp|P13674|P4HA1_HUMAN **TwLFY**
sp|Q6IFH4|OR6B2_HUMAN **TqLYF**
sp|P15586|GNS_HUMAN **TyIFY**
sp|P30519|HMOX2_HUMAN **TaLYF**
sp|P51690|ARSE_HUMAN **TlIYF**
sp|Q8NGG8|OR8B3_HUMAN **TqLFF**
sp|Q6ZPD9|D19L3_HUMAN **TlLFY**
sp|A3KFT3|OR2M5_HUMAN **TfLFF**
sp|A3KFT3|OR2M5_HUMAN_2 **TqIFF**
sp|Q8IYU2|HACE1_HUMAN **TpLYF**

sp|Q8N0Y3|OR4N4_HUMAN **TqLFF**
sp|P09758|TACD2_HUMAN **TlIYY**
sp|P32302|CXCR5_HUMAN **TsIFF**
sp|Q9UM44|HHLA2_HUMAN **TsLFY**
sp|P48066|S6A11_HUMAN **TqIFF**

8.3. Appendix 3

Excel file of functional enrichment analysis of motif hits list in Appendix 2 using FunRich.

8.4. Appendix 4

Raw graphical data summarising effects of the S134A mutation on the overall conformation of AGR2 protein analysed by hydrogen-deuterium exchange mass spectrometry.

8.5. Appendix 5

Excel file of TMT mass spectrometry quantitation of FLO-1 cells with transient overexpression of wt-AGR2, AGR2-KDEL, or EV control.

CHAPTER 9: References

1. Pecorino, L., *Molecular biology of cancer : mechanisms, targets, and therapeutics*. 2016, New York, NY: Oxford University Press. pages cm.
2. Paget, S., *The distribution of secondary growths in cancer of the breast. 1889.* Cancer Metastasis Rev, 1989. **8**(2): p. 98-101.
3. Fidler, I.J., *The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited.* Nat Rev Cancer, 2003. **3**(6): p. 453-8.
4. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
5. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
6. Scott, A.M., et al., *A phase I clinical trial with monoclonal antibody ch806 targeting transitional state and mutant epidermal growth factor receptors.* Proc Natl Acad Sci U S A, 2007. **104**(10): p. 4071-6.
7. Urruticoechea, A., et al., *Recent advances in cancer therapy: an overview.* Curr Pharm Des, 2010. **16**(1): p. 3-10.
8. Scott, A.M., J.D. Wolchok, and L.J. Old, *Antibody therapy of cancer.* Nat Rev Cancer, 2012. **12**(4): p. 278-87.
9. Vanneman, M. and G. Dranoff, *Combining immunotherapy and targeted therapies in cancer treatment.* Nat Rev Cancer, 2012. **12**(4): p. 237-51.
10. Langley, R.R. and I.J. Fidler, *Tumor cell-organ microenvironment interactions in the pathogenesis of cancer metastasis.* Endocr Rev, 2007. **28**(3): p. 297-321.
11. Talmadge, J.E., et al., *Development of biological diversity and susceptibility to chemotherapy in murine cancer metastases.* Cancer Res, 1984. **44**(9): p. 3801-5.
12. Spang, A., *On vesicle formation and tethering in the ER-Golgi shuttle.* Curr Opin Cell Biol, 2009. **21**(4): p. 531-6.
13. Mellman, I. and G. Warren, *The road taken: past and future foundations of membrane traffic.* Cell, 2000. **100**(1): p. 99-112.
14. Ghaemmaghami, S., et al., *Global analysis of protein expression in yeast.* Nature, 2003. **425**(6959): p. 737-41.
15. Ellgaard, L. and A. Helenius, *Quality control in the endoplasmic reticulum.* Nat Rev Mol Cell Biol, 2003. **4**(3): p. 181-91.
16. Vitale, A. and J. Denecke, *The endoplasmic reticulum-gateway of the secretory pathway.* Plant Cell, 1999. **11**(4): p. 615-28.
17. Ellis, R.J., *Molecular chaperones: assisting assembly in addition to folding.* Trends Biochem Sci, 2006. **31**(7): p. 395-401.
18. Miller, E.A., et al., *Multiple cargo binding sites on the COPII subunit Sec24p ensure capture of diverse membrane proteins into transport vesicles.* Cell, 2003. **114**(4): p. 497-509.
19. Xu, D. and J.C. Hay, *Reconstitution of COPII vesicle fusion to generate a pre-Golgi intermediate compartment.* J Cell Biol, 2004. **167**(6): p. 997-1003.
20. Appenzeller-Herzog, C. and H.P. Hauri, *The ER-Golgi intermediate compartment (ERGIC): in search of its identity and function.* J Cell Sci, 2006. **119**(Pt 11): p. 2173-83.

21. Pelham, H.R., *SNAREs and the specificity of membrane fusion.* Trends Cell Biol, 2001. **11**(3): p. 99-101.
22. Jackson, C.L., *Mechanisms of transport through the Golgi complex.* J Cell Sci, 2009. **122**(Pt 4): p. 443-52.
23. Rabouille, C. and J. Klumperman, *Opinion: The maturing role of COPI vesicles in intra-Golgi transport.* Nat Rev Mol Cell Biol, 2005. **6**(10): p. 812-7.
24. Jahn, R. and R.H. Scheller, *SNAREs--engines for membrane fusion.* Nat Rev Mol Cell Biol, 2006. **7**(9): p. 631-43.
25. Nickel, W. and M. Seedorf, *Unconventional mechanisms of protein transport to the cell surface of eukaryotic cells.* Annu Rev Cell Dev Biol, 2008. **24**: p. 287-308.
26. Fatal, N., et al., *Active and specific recruitment of a soluble cargo protein for endoplasmic reticulum exit in the absence of functional COPII component Sec24p.* J Cell Sci, 2004. **117**(Pt 9): p. 1665-73.
27. Yoo, J.S., et al., *Non-conventional trafficking of the cystic fibrosis transmembrane conductance regulator through the early secretory pathway.* J Biol Chem, 2002. **277**(13): p. 11401-9.
28. Nickel, W. and C. Rabouille, *Mechanisms of regulated unconventional protein secretion.* Nat Rev Mol Cell Biol, 2009. **10**(2): p. 148-55.
29. Hetz, C. and L.H. Glimcher, *Protein homeostasis networks in physiology and disease.* Curr Opin Cell Biol, 2011. **23**(2): p. 123-5.
30. Powers, E.T. and W.E. Balch, *Diversity in the origins of proteostasis networks--a driver for protein function in evolution.* Nat Rev Mol Cell Biol, 2013. **14**(4): p. 237-48.
31. Schroder, M. and R.J. Kaufman, *The mammalian unfolded protein response.* Annu Rev Biochem, 2005. **74**: p. 739-89.
32. Woehlbier, U. and C. Hetz, *Modulating stress responses by the UPRosome: a matter of life and death.* Trends Biochem Sci, 2011. **36**(6): p. 329-37.
33. Hetz, C., et al., *The unfolded protein response: integrating stress signals through the stress sensor IRE1alpha.* Physiol Rev, 2011. **91**(4): p. 1219-43.
34. Hetz, C., E. Chevet, and S.A. Oakes, *Proteostasis control by the unfolded protein response.* Nat Cell Biol, 2015. **17**(7): p. 829-38.
35. Ron, D. and P. Walter, *Signal integration in the endoplasmic reticulum unfolded protein response.* Nat Rev Mol Cell Biol, 2007. **8**(7): p. 519-29.
36. Tabas, I. and D. Ron, *Integrating the mechanisms of apoptosis induced by endoplasmic reticulum stress.* Nat Cell Biol, 2011. **13**(3): p. 184-90.
37. Li, Z. and Z. Li, *Glucose regulated protein 78: a critical link between tumor microenvironment and cancer hallmarks.* Biochim Biophys Acta, 2012. **1826**(1): p. 13-22.
38. Moenner, M., et al., *Integrated endoplasmic reticulum stress responses in cancer.* Cancer Res, 2007. **67**(22): p. 10631-4.
39. Dejeans, N., et al., *Addicted to secrete - novel concepts and targets in cancer therapy.* Trends Mol Med, 2014. **20**(5): p. 242-50.
40. Dejeans, N., et al., *Novel roles of the unfolded protein response in the control of tumor development and aggressiveness.* Semin Cancer Biol, 2015. **33**: p. 67-73.

41. Hetz, C., E. Chevet, and H.P. Harding, *Targeting the unfolded protein response in disease.* Nat Rev Drug Discov, 2013. **12**(9): p. 703-19.
42. Ali, M.M., et al., *Structure of the Ire1 autophosphorylation complex and implications for the unfolded protein response.* EMBO J, 2011. **30**(5): p. 894-905.
43. Axten, J.M., et al., *Discovery of 7-methyl-5-(1-{[3-(trifluoromethyl)phenyl]acetyl}-2,3-dihydro-1H-indol-5-yl)-7H-p yrrolo[2,3-d]pyrimidin-4-amine (GSK2606414), a potent and selective first-in-class inhibitor of protein kinase R (PKR)-like endoplasmic reticulum kinase (PERK).* J Med Chem, 2012. **55**(16): p. 7193-207.
44. Schubert, U., et al., *Rapid degradation of a large fraction of newly synthesized proteins by proteasomes.* Nature, 2000. **404**(6779): p. 770-4.
45. Jahn, T.R. and S.E. Radford, *The Yin and Yang of protein folding.* FEBS J, 2005. **272**(23): p. 5962-70.
46. McCracken, A.A. and J.L. Brodsky, *Assembly of ER-associated protein degradation in vitro: dependence on cytosol, calnexin, and ATP.* J Cell Biol, 1996. **132**(3): p. 291-8.
47. Vembar, S.S. and J.L. Brodsky, *One step at a time: endoplasmic reticulum-associated degradation.* Nat Rev Mol Cell Biol, 2008. **9**(12): p. 944-57.
48. Ng, D.T., E.D. Spear, and P. Walter, *The unfolded protein response regulates multiple aspects of secretory and membrane protein biogenesis and endoplasmic reticulum quality control.* J Cell Biol, 2000. **150**(1): p. 77-88.
49. Travers, K.J., et al., *Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation.* Cell, 2000. **101**(3): p. 249-58.
50. Nair, U. and D.J. Klionsky, *Molecular mechanisms and regulation of specific and nonspecific autophagy pathways in yeast.* J Biol Chem, 2005. **280**(51): p. 41785-8.
51. Sive, H.L., K. Hattori, and H. Weintraub, *Progressive determination during formation of the anteroposterior axis in Xenopus laevis.* Cell, 1989. **58**(1): p. 171-80.
52. Aberger, F., et al., *Anterior specification of embryonic ectoderm: the role of the Xenopus cement gland-specific gene XAG-2.* Mech Dev, 1998. **72**(1-2): p. 115-30.
53. Novoselov, V.V., et al., *Expression zones of three novel genes abut the developing anterior neural plate of Xenopus embryo.* Gene Expr Patterns, 2003. **3**(2): p. 225-30.
54. Kumar, A., et al., *Molecular basis for the nerve dependence of limb regeneration in an adult vertebrate.* Science, 2007. **318**(5851): p. 772-7.
55. da Silva, S.M., P.B. Gates, and J.P. Brockes, *The newt ortholog of CD59 is implicated in proximodistal identity during amphibian limb regeneration.* Dev Cell, 2002. **3**(4): p. 547-55.
56. Blassberg, R.A., et al., *Functional convergence of signalling by GPI-anchored and anchorless forms of a salamander protein implicated in limb regeneration.* J Cell Sci, 2011. **124**(Pt 1): p. 47-56.
57. Ivanova, A.S., et al., *Agr genes, missing in amniotes, are involved in the body appendages regeneration in frog tadpoles.* Sci Rep, 2013. **3**: p. 1279.

58. Fletcher, G.C., et al., *hAG-2 and hAG-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumours and interact with metastasis gene C4.4a and dystroglycan.* Br J Cancer, 2003. **88**(4): p. 579-85.
59. Komiya, T., Y. Tanigawa, and S. Hirohashi, *Cloning of the gene gob-4, which is expressed in intestinal goblet cells in mice.* Biochim Biophys Acta, 1999. **1444**(3): p. 434-8.
60. Park, S.W., et al., *The protein disulfide isomerase AGR2 is essential for production of intestinal mucus.* Proc Natl Acad Sci U S A, 2009. **106**(17): p. 6950-5.
61. Patel, P., et al., *Metastasis-promoting anterior gradient 2 protein has a dimeric thioredoxin fold structure and a role in cell adhesion.* J Mol Biol, 2013. **425**(5): p. 929-43.
62. Wang, Z., Y. Hao, and A.W. Lowe, *The adenocarcinoma-associated antigen, AGR2, promotes tumor growth, cell migration, and cellular transformation.* Cancer Res, 2008. **68**(2): p. 492-7.
63. Zhao, F., et al., *Disruption of Paneth and goblet cell homeostasis and increased endoplasmic reticulum stress in Agr2-/- mice.* Dev Biol, 2010. **338**(2): p. 270-9.
64. Ganz, T., *Paneth cells--guardians of the gut cell hatchery.* Nat Immunol, 2000. **1**(2): p. 99-100.
65. Keshav, S., et al., *Tumor necrosis factor mRNA localized to Paneth cells of normal murine intestinal epithelium by in situ hybridization.* J Exp Med, 1990. **171**(1): p. 327-32.
66. Persson, S., et al., *Diversity of the protein disulfide isomerase family: identification of breast tumor induced Hag2 and Hag3 as novel members of the protein family.* Mol Phylogenet Evol, 2005. **36**(3): p. 734-40.
67. Ivanova, A.S., et al., *The secreted factor Ag1 missing in higher vertebrates regulates fins regeneration in Danio rerio.* Sci Rep, 2015. **5**: p. 8123.
68. Grassme, K.S., et al., *Mechanism of Action of Secreted Newt Anterior Gradient Protein.* PLoS One, 2016. **11**(4): p. e0154176.
69. Petek, E., et al., *Localization of the human anterior gradient-2 gene (AGR2) to chromosome band 7p21.3 by radiation hybrid mapping and fluorescencein situ hybridisation.* Cytogenet Cell Genet, 2000. **89**(3-4): p. 141-2.
70. Zhang, J.S., et al., *AGR2, an androgen-inducible secretory protein overexpressed in prostate cancer.* Genes Chromosomes Cancer, 2005. **43**(3): p. 249-59.
71. Zheng, W., et al., *Evaluation of AGR2 and AGR3 as candidate genes for inflammatory bowel disease.* Genes Immun, 2006. **7**(1): p. 11-8.
72. Krig, S.R., et al., *Identification of genes directly regulated by the oncogene ZNF217 using chromatin immunoprecipitation (ChIP)-chip assays.* J Biol Chem, 2007. **282**(13): p. 9703-12.
73. Wilson, C.L., et al., *Effects of oestrogen on gene expression in epithelium and stroma of normal human breast tissue.* Endocr Relat Cancer, 2006. **13**(2): p. 617-28.
74. Zweitzig, D.R., et al., *Physiological stress induces the metastasis marker AGR2 in breast cancer cells.* Mol Cell Biochem, 2007. **306**(1-2): p. 255-60.

75. Higa, A., et al., *Role of pro-oncogenic protein disulfide isomerase (PDI) family member anterior gradient 2 (AGR2) in the control of endoplasmic reticulum homeostasis.* J Biol Chem, 2011. **286**(52): p. 44855-68.
76. Norris, A.M., et al., *AGR2 is a SMAD4-suppressible gene that modulates MUC1 levels and promotes the initiation and progression of pancreatic intraepithelial neoplasia.* Oncogene, 2013. **32**(33): p. 3867-76.
77. Myung, J.K., et al., *Mass spectrometrical analysis of the processed metastasis-inducing anterior gradient protein 2 homolog reveals 100% sequence coverage.* Amino Acids, 2008. **35**(2): p. 485-94.
78. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity.* Bioinformatics, 2015. **31**(6): p. 857-63.
79. Iakoucheva, L.M., et al., *Intrinsic disorder in cell-signaling and cancer-associated proteins.* J Mol Biol, 2002. **323**(3): p. 573-84.
80. Voss, M., B. Schroder, and R. Fluhrer, *Mechanism, specificity, and physiology of signal peptide peptidase (SPP) and SPP-like proteases.* Biochim Biophys Acta, 2013. **1828**(12): p. 2828-39.
81. Fourtouna, A., et al., *The anterior gradient-2 pathway as a model for developing peptide-aptamer anti-cancer dtug leads that stimulate p53 function.* current chemical biology, 2009. **3**: p. 124-137.
82. Chen, C.Y., et al., *Signal peptide peptidase functions in ERAD to cleave the unfolded protein response regulator XBP1u.* EMBO J, 2014. **33**(21): p. 2492-506.
83. Gupta, A., A. Dong, and A.W. Lowe, *AGR2 gene function requires a unique endoplasmic reticulum localization motif.* J Biol Chem, 2012. **287**(7): p. 4773-82.
84. Munro, S. and H.R. Pelham, *A C-terminal signal prevents secretion of luminal ER proteins.* Cell, 1987. **48**(5): p. 899-907.
85. Alanen, H.I., et al., *Beyond KDEL: the role of positions 5 and 6 in determining ER localization.* J Mol Biol, 2011. **409**(3): p. 291-7.
86. Raykhel, I., et al., *A molecular specificity code for the three mammalian KDEL receptors.* J Cell Biol, 2007. **179**(6): p. 1193-204.
87. Pohler, E., et al., *The Barrett's antigen anterior gradient-2 silences the p53 transcriptional response to DNA damage.* Mol Cell Proteomics, 2004. **3**(6): p. 534-47.
88. Wunderlich, M., et al., *Efficient catalysis of disulfide formation during protein folding with a single active-site cysteine.* J Mol Biol, 1995. **247**(1): p. 28-33.
89. Norgaard, P. and J.R. Winther, *Mutation of yeast Eug1p CXXS active sites to CXXC results in a dramatic increase in protein disulphide isomerase activity.* Biochem J, 2001. **358**(Pt 1): p. 269-74.
90. Anelli, T., et al., *Thiol-mediated protein retention in the endoplasmic reticulum: the role of ERp44.* EMBO J, 2003. **22**(19): p. 5015-22.
91. Anelli, T., et al., *ERp44, a novel endoplasmic reticulum folding assistant of the thioredoxin family.* EMBO J, 2002. **21**(4): p. 835-44.
92. Rowe, M.L., et al., *Solution structure and dynamics of ERp18, a small endoplasmic reticulum resident oxidoreductase.* Biochemistry, 2009. **48**(21): p. 4596-606.

93. Alanen, H.I., et al., *Functional characterization of ERp18, a new endoplasmic reticulum-located thioredoxin superfamily member.* J Biol Chem, 2003. **278**(31): p. 28912-20.
94. Chevet, E., et al., *Emerging roles for the pro-oncogenic anterior gradient-2 in cancer development.* Oncogene, 2013. **32**(20): p. 2499-509.
95. Bonser, L.R., et al., *The Endoplasmic Reticulum Resident Protein AGR3. Required for Regulation of Ciliary Beat Frequency in the Airway.* Am J Respir Cell Mol Biol, 2015. **53**(4): p. 536-43.
96. Bergstrom, J.H., et al., *AGR2, an endoplasmic reticulum protein, is secreted into the gastrointestinal mucus.* PLoS One, 2014. **9**(8): p. e104186.
97. Gray, T.A., et al., *Development of a fluorescent monoclonal antibody-based assay to measure the allosteric effects of synthetic peptides on self-oligomerization of AGR2 protein.* Protein Sci, 2013. **22**(9): p. 1266-78.
98. Ryu, J., et al., *Dimerization of pro-oncogenic protein Anterior Gradient 2 is required for the interaction with BiP/GRP78.* Biochem Biophys Res Commun, 2013. **430**(2): p. 610-5.
99. Clarke, D.J., et al., *Mass spectrometry analysis of the oxidation states of the pro-oncogenic protein anterior gradient-2 reveals covalent dimerization via an intermolecular disulphide bond.* Biochim Biophys Acta, 2016. **1864**(5): p. 551-61.
100. Dumartin, L., et al., *AGR2 is a novel surface antigen that promotes the dissemination of pancreatic cancer cells through regulation of cathepsins B and D.* Cancer Res, 2011. **71**(22): p. 7091-102.
101. Brychtova, V., B. Vojtesek, and R. Hrstka, *Anterior gradient 2: a novel player in tumor cell biology.* Cancer Lett, 2011. **304**(1): p. 1-7.
102. Xia, J.H., et al., *Predominant expression and cellular distribution of fish Agr2 in renal collecting system.* Comp Biochem Physiol B Biochem Mol Biol, 2009. **152**(4): p. 397-404.
103. Verma, S., et al., *The estrogen-responsive Agr2 gene regulates mammary epithelial proliferation and facilitates lobuloalveolar development.* Dev Biol, 2012. **369**(2): p. 249-60.
104. Schroeder, B.W., et al., *AGR2 is induced in asthma and promotes allergen-induced mucin overproduction.* Am J Respir Cell Mol Biol, 2012. **47**(2): p. 178-85.
105. Maslon, M.M., et al., *A divergent substrate-binding loop within the pro-oncogenic protein anterior gradient-2 forms a docking site for Reptin.* J Mol Biol, 2010. **404**(3): p. 418-38.
106. Nano, N. and W.A. Houry, *Chaperone-like activity of the AAA+ proteins Rvb1 and Rvb2 in the assembly of various complexes.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1617): p. 20110399.
107. Zhao, L., et al., *Reptin/Ruvbl2 is a Lrrc6/Seahorse interactor essential for cilia motility.* Proc Natl Acad Sci U S A, 2013. **110**(31): p. 12697-702.
108. Zhao, T.J., et al., *Catalysis of creatine kinase refolding by protein disulfide isomerase involves disulfide cross-link and dimer to tetramer switch.* J Biol Chem, 2005. **280**(14): p. 13470-6.
109. Berggard, T., S. Linse, and P. James, *Methods for the detection and analysis of protein-protein interactions.* Proteomics, 2007. **7**(16): p. 2833-42.

110. Phizicky, E.M. and S. Fields, *Protein-protein interactions: methods for detection and analysis.* Microbiol Rev, 1995. **59**(1): p. 94-123.
111. Dong, A., D. Wodziak, and A.W. Lowe, *Epidermal growth factor receptor (EGFR) signaling requires a specific endoplasmic reticulum thioredoxin for the post-translational control of receptor presentation to the cell surface.* J Biol Chem, 2015. **290**(13): p. 8016-27.
112. Hatahet, F. and L.W. Ruddock, *Protein disulfide isomerase: a critical evaluation of its function in disulfide bond formation.* Antioxid Redox Signal, 2009. **11**(11): p. 2807-50.
113. Laurindo, F.R., L.A. Pescatore, and C. Fernandes Dde, *Protein disulfide isomerase in redox cell signaling and homeostasis.* Free Radic Biol Med, 2012. **52**(9): p. 1954-69.
114. Galligan, J.J. and D.R. Petersen, *The human protein disulfide isomerase gene family.* Hum Genomics, 2012. **6**: p. 6.
115. Benham, A.M., *The protein disulfide isomerase family: key players in health and disease.* Antioxid Redox Signal, 2012. **16**(8): p. 781-9.
116. Brychtova, V., et al., *Mechanisms of anterior gradient-2 regulation and function in cancer.* Semin Cancer Biol, 2015. **33**: p. 16-24.
117. Di Valentin, E., et al., *New asthma biomarkers: lessons from murine models of acute and chronic asthma.* Am J Physiol Lung Cell Mol Physiol, 2009. **296**(2): p. L185-97.
118. Yu, H., et al., *Interleukin-13 induces mucin 5AC production involving STAT6/SPDEF in human airway epithelial cells.* Cell Commun Adhes, 2010. **17**(4-6): p. 83-92.
119. Kuyper, L.M., et al., *Characterization of airway plugging in fatal asthma.* Am J Med, 2003. **115**(1): p. 6-11.
120. Ordonez, C.L., et al., *Mild and moderate asthma is associated with airway goblet cell hyperplasia and abnormalities in mucin gene expression.* Am J Respir Crit Care Med, 2001. **163**(2): p. 517-23.
121. Zhou, M., et al., *Effect of dexamethasone on expression of AGR2 protein in asthmatic mice.* J Huazhong Univ Sci Technolog Med Sci, 2013. **33**(1): p. 33-6.
122. Satsangi, J., et al., *Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12.* Nat Genet, 1996. **14**(2): p. 199-202.
123. Hampe, J., et al., *A genomewide analysis provides evidence for novel linkages in inflammatory bowel disease in a large European cohort.* Am J Hum Genet, 1999. **64**(3): p. 808-16.
124. Bogaert, S., et al., *Involvement of endoplasmic reticulum stress in inflammatory bowel disease: a different implication for colonic and ileal disease?* PLoS One, 2011. **6**(10): p. e25589.
125. Niederreiter, L. and A. Kaser, *Endoplasmic reticulum stress and inflammatory bowel disease.* Acta Gastroenterol Belg, 2011. **74**(2): p. 330-3.
126. Thompson, D.A. and R.J. Weigel, *hAG-2, the human homologue of the Xenopus laevis cement gland gene XAG-2, is coexpressed with estrogen receptor in breast cancer cell lines.* Biochem Biophys Res Commun, 1998. **251**(1): p. 111-6.

127. Hrstka, R., et al., *The pro-metastatic protein anterior gradient-2 predicts poor prognosis in tamoxifen-treated breast cancers.* Oncogene, 2010. **29**(34): p. 4838-47.
128. Adam, P.J., et al., *Comprehensive proteomic analysis of breast cancer cell membranes reveals unique proteins with potential roles in clinical cancer.* J Biol Chem, 2003. **278**(8): p. 6482-9.
129. Liu, D., et al., *Human homologue of cement gland protein, a novel metastasis inducer associated with breast carcinomas.* Cancer Res, 2005. **65**(9): p. 3796-805.
130. Barraclough, D.L., et al., *The metastasis-associated anterior gradient 2 protein is correlated with poor survival of breast cancer patients.* Am J Pathol, 2009. **175**(5): p. 1848-57.
131. Fritzsche, F.R., et al., *Prognostic relevance of AGR2 expression in breast cancer.* Clin Cancer Res, 2006. **12**(6): p. 1728-34.
132. Wu, Z.S., et al., *[Expression of a novel metastasis-inducing protein human anterior gradient-2 (AGR2) in breast cancer and its clinical and prognostic significance].* Zhonghua Bing Li Xue Za Zhi, 2008. **37**(2): p. 109-13.
133. Bu, H., et al., *Anterior gradient 2 and 3--two prototype androgen-responsive genes transcriptionally upregulated by androgens and by oestrogens in prostate cancer cells.* FEBS J, 2013. **280**(5): p. 1249-66.
134. Zhang, Y., et al., *Increased expression of anterior gradient-2 is significantly associated with poor survival of prostate cancer patients.* Prostate Cancer Prostatic Dis, 2007. **10**(3): p. 293-300.
135. Edgell, T.A., et al., *Increased plasma concentrations of anterior gradient 2 protein are positively associated with ovarian cancer.* Clin Sci (Lond), 2010. **118**(12): p. 717-25.
136. Gray, T.A., et al., *Anterior Gradient-3: a novel biomarker for ovarian cancer that mediates cisplatin resistance in xenograft models.* J Immunol Methods, 2012. **378**(1-2): p. 20-32.
137. Park, K., et al., *AGR2, a mucinous ovarian cancer marker, promotes cell proliferation and migration.* Exp Mol Med, 2011. **43**(2): p. 91-100.
138. Rice, G.E., T.A. Edgell, and D.J. Autelitano, *Evaluation of midkine and anterior gradient 2 in a multimarker panel for the detection of ovarian cancer.* J Exp Clin Cancer Res, 2010. **29**: p. 62.
139. Yagui-Beltran, A., et al., *The human oesophageal squamous epithelium exhibits a novel type of heat shock protein response.* Eur J Biochem, 2001. **268**(20): p. 5343-55.
140. Lee do, H., et al., *Identification of proteins differentially expressed in gastric cancer cells with high metastatic potential for invasion to lymph nodes.* Mol Cells, 2011. **31**(6): p. 563-71.
141. Rusch, V., et al., *Overexpression of the epidermal growth factor receptor and its ligand transforming growth factor alpha is frequent in resectable non-small cell lung cancer but does not predict tumor progression.* Clin Cancer Res, 1997. **3**(4): p. 515-22.
142. Fritzsche, F.R., et al., *Expression of AGR2 in non small cell lung cancer.* Histol Histopathol, 2007. **22**(7): p. 703-8.

143. Ramachandran, V., et al., *Anterior gradient 2 is expressed and secreted during the development of pancreatic cancer and promotes cancer cell survival.* Cancer Res, 2008. **68**(19): p. 7811-8.
144. Vanderlaag, K.E., et al., *Anterior gradient-2 plays a critical role in breast cancer cell growth and survival by modulating cyclin D1, estrogen receptor-alpha and survivin.* Breast Cancer Res, 2010. **12**(3): p. R32.
145. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.
146. Salmans, M.L., F. Zhao, and B. Andersen, *The estrogen-regulated anterior gradient 2 (AGR2) protein in breast cancer: a potential drug target and biomarker.* Breast Cancer Res, 2013. **15**(2): p. 204.
147. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
148. Armes, J.E., et al., *AGR2 expression in ovarian tumours: a potential biomarker for endometrioid and mucinous differentiation.* Pathology, 2013. **45**(1): p. 49-54.
149. Darb-Esfahani, S., et al., *Anterior gradient protein 2 (AGR2) is an independent prognostic factor in ovarian high-grade serous carcinoma.* Virchows Arch, 2012. **461**(2): p. 109-16.
150. Kristiansen, G., et al., *Expression profiling of microdissected matched prostate cancer samples reveals CD166/MEMD and CD24 as new prognostic markers for patient survival.* J Pathol, 2005. **205**(3): p. 359-76.
151. Kani, K., et al., *Anterior gradient 2 (AGR2): blood-based biomarker elevated in metastatic prostate cancer associated with the neuroendocrine phenotype.* Prostate, 2013. **73**(3): p. 306-15.
152. Bu, H., et al., *The anterior gradient 2 (AGR2) gene is overexpressed in prostate cancer and may be useful as a urine sediment marker for prostate cancer detection.* Prostate, 2011. **71**(6): p. 575-87.
153. Lepreux, S., P. Bioulac-Sage, and E. Chevet, *Differential expression of the anterior gradient protein-2 is a conserved feature during morphogenesis and carcinogenesis of the biliary tree.* Liver Int, 2011. **31**(3): p. 322-8.
154. Varghese, S., P. Lao-Sirieix, and R.C. Fitzgerald, *Identification and clinical implementation of biomarkers for Barrett's esophagus.* Gastroenterology, 2012. **142**(3): p. 435-441 e2.
155. Valladares-Ayerbes, M., et al., *Evaluation of the adenocarcinoma-associated gene AGR2 and the intestinal stem cell marker LGR5 as biomarkers in colorectal cancer.* Int J Mol Sci, 2012. **13**(4): p. 4367-87.
156. Chung, K., et al., *Serum AGR2 as an early diagnostic and postoperative prognostic biomarker of human lung adenocarcinoma.* Cancer Biomark, 2011. **10**(2): p. 101-7.
157. Sweeny, L., et al., *CD147 and AGR2 expression promote cellular proliferation and metastasis of head and neck squamous cell carcinoma.* Exp Cell Res, 2012. **318**(14): p. 1788-98.
158. Li, Z., et al., *Induction of anterior gradient 2 (AGR2) plays a key role in insulin-like growth factor-1 (IGF-1)-induced breast cancer cell proliferation and migration.* Med Oncol, 2015. **32**(6): p. 577.

159. Duran, M.C., et al., *Characterisation of tumoral markers correlated with ErbB2 (HER2/Neu) overexpression and metastasis in breast cancer.* Proteomics Clin Appl, 2008. **2**(9): p. 1313-26.
160. Zhang, Y., et al., *ErbB3 binding protein 1 represses metastasis-promoting gene anterior gradient protein 2 in prostate cancer.* Cancer Res, 2010. **70**(1): p. 240-8.
161. Hu, Z., et al., *Knockdown of AGR2 induces cellular senescence in prostate cancer cells.* Carcinogenesis, 2012. **33**(6): p. 1178-86.
162. Barry, S., et al., *S100P is a metastasis-associated gene that facilitates transendothelial migration of pancreatic cancer cells.* Clin Exp Metastasis, 2013. **30**(3): p. 251-64.
163. Chen, R., et al., *Elevated level of anterior gradient-2 in pancreatic juice from patients with pre-malignant pancreatic neoplasia.* Mol Cancer, 2010. **9**: p. 149.
164. Makawita, S., et al., *Integrated proteomic profiling of cell line conditioned media and pancreatic juice for the identification of pancreatic cancer biomarkers.* Mol Cell Proteomics, 2011. **10**(10): p. M111 008599.
165. Vitello, E.A., et al., *Cancer-secreted AGR2 induces programmed cell death in normal cells.* Oncotarget, 2016.
166. Missiaglia, E., et al., *Analysis of gene expression in cancer cell lines identifies candidate markers for pancreatic tumorigenesis and metastasis.* Int J Cancer, 2004. **112**(1): p. 100-12.
167. Mani, S.A., et al., *The epithelial-mesenchymal transition generates cells with properties of stem cells.* Cell, 2008. **133**(4): p. 704-15.
168. Thiery, J.P., *Epithelial-mesenchymal transitions in development and pathologies.* Curr Opin Cell Biol, 2003. **15**(6): p. 740-6.
169. Dong, A., et al., *The human adenocarcinoma-associated gene, AGR2, induces expression of amphiregulin through Hippo pathway co-activator YAP1 activation.* J Biol Chem, 2011. **286**(20): p. 18301-10.
170. Zeng, Q. and W. Hong, *The emerging role of the hippo pathway in cell contact inhibition, organ size control, and cancer development in mammals.* Cancer Cell, 2008. **13**(3): p. 188-92.
171. Vendrell, J.A., et al., *A candidate molecular signature associated with tamoxifen failure in primary breast cancer.* Breast Cancer Res, 2008. **10**(5): p. R88.
172. Hrstka, R., et al., *AGR2 predicts tamoxifen resistance in postmenopausal breast cancer patients.* Dis Markers, 2013. **35**(4): p. 207-12.
173. Hengel, S.M., et al., *Data-independent proteomic screen identifies novel tamoxifen agonist that mediates drug resistance.* J Proteome Res, 2011. **10**(10): p. 4567-78.
174. Hrstka, R., et al., *Identification of an AKT-dependent signalling pathway that mediates tamoxifen-dependent induction of the pro-metastatic protein anterior gradient-2.* Cancer Lett, 2013. **333**(2): p. 187-93.
175. Huber, M., et al., *Comparison of proteomic and genomic analyses of the human breast cancer cell line T47D and the antiestrogen-resistant derivative T47D-r.* Mol Cell Proteomics, 2004. **3**(1): p. 43-55.
176. Zhao, L., et al., *Identification of candidate biomarkers of therapeutic response to docetaxel by proteomic profiling.* Cancer Res, 2009. **69**(19): p. 7696-703.

177. Kang, K., et al., *A role for protein disulfide isomerase in the early folding and assembly of MHC class I molecules.* Antioxid Redox Signal, 2009. **11**(10): p. 2553-61.
178. Luo, B. and A.S. Lee, *The critical roles of endoplasmic reticulum chaperones and unfolded protein response in tumorigenesis and anticancer therapies.* Oncogene, 2013. **32**(7): p. 805-18.
179. Turano, C., et al., *Proteins of the PDI family: unpredicted non-ER locations and functions.* J Cell Physiol, 2002. **193**(2): p. 154-63.
180. Fessart, D., et al., *Secretion of protein disulphide isomerase AGR2 confers tumorigenic properties.* Elife, 2016. **5**.
181. Fourtouna, A., et al., *The Anterior Gradient-2 Pathway as a Model for Developing Peptide-Aptamer Anti-Cancer Drug Leads that Stimulate p53 Function.* Curr Chem Biol, 2009. **3**(2): p. 124-137.
182. Hernychova, L., et al., *Identification of a second Nutlin-3 responsive interaction site in the N-terminal domain of MDM2 using hydrogen/deuterium exchange mass spectrometry.* Proteomics, 2013. **13**(16): p. 2512-25.
183. Kavan, D. and P. Man, *MSTools - Web based application for visualization and presentation of HXMS data.* International Journal of Mass Spectrometry, 2011. **302**: p. 53-58.
184. Wisniewski, J.R., et al., *Universal sample preparation method for proteome analysis.* Nat Methods, 2009. **6**(5): p. 359-62.
185. Hengel, S.M., et al., *Data-independent proteomic screen identifies novel tamoxifen agonist that mediates drug resistance.* Journal of proteome research, 2011. **10**(10): p. 4567-78.
186. Gray, T.A., et al., *Anterior Gradient-3: a novel biomarker for ovarian cancer that mediates cisplatin resistance in xenograft models.* Journal of immunological methods, 2012. **378**(1-2): p. 20-32.
187. Dong, A., et al., *The human adenocarcinoma-associated gene, AGR2, induces expression of amphiregulin through Hippo pathway co-activator YAP1 activation.* The Journal of biological chemistry, 2011. **286**(20): p. 18301-10.
188. Sammut, S.J., R.D. Finn, and A. Bateman, *Pfam 10 years on: 10,000 families and still growing.* Brief Bioinform, 2008. **9**(3): p. 210-9.
189. Aloy, P. and R.B. Russell, *Ten thousand interactions for the molecular biologist.* Nat Biotechnol, 2004. **22**(10): p. 1317-21.
190. Hart, G.T., A.K. Ramani, and E.M. Marcotte, *How complete are current yeast and human protein-interaction networks?* Genome Biol, 2006. **7**(11): p. 120.
191. Stumpf, M.P., et al., *Estimating the size of the human interactome.* Proc Natl Acad Sci U S A, 2008. **105**(19): p. 6959-64.
192. Venkatesan, K., et al., *An empirical framework for binary interactome mapping.* Nat Methods, 2009. **6**(1): p. 83-90.
193. Havugimana, P.C., et al., *A census of human soluble protein complexes.* Cell, 2012. **150**(5): p. 1068-81.
194. Levy, E.D., C.R. Landry, and S.W. Michnick, *How perfect can protein interactomes be?* Sci Signal, 2009. **2**(60): p. pe11.
195. Perica, T., et al., *The emergence of protein complexes: quaternary structure, dynamics and allostery. Colworth Medal Lecture.* Biochem Soc Trans, 2012. **40**(3): p. 475-91.

196. Tompa, P., et al., *A million peptide motifs for the molecular biologist.* Mol Cell, 2014. **55**(2): p. 161-9.
197. Glotzer, M., A.W. Murray, and M.W. Kirschner, *Cyclin is degraded by the ubiquitin pathway.* Nature, 1991. **349**(6305): p. 132-8.
198. Puntervoll, P., et al., *ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins.* Nucleic Acids Res, 2003. **31**(13): p. 3625-30.
199. Perkins, J.R., et al., *Transient protein-protein interactions: structural, functional, and network properties.* Structure, 2010. **18**(10): p. 1233-43.
200. Van Roey, K., T.J. Gibson, and N.E. Davey, *Motif switches: decision-making in cell regulation.* Curr Opin Struct Biol, 2012. **22**(3): p. 378-85.
201. Stein, A., et al., *Dynamic interactions of proteins in complex networks: a more structured view.* FEBS J, 2009. **276**(19): p. 5390-405.
202. Diella, F., et al., *Understanding eukaryotic linear motifs and their role in cell signaling and regulation.* Front Biosci, 2008. **13**: p. 6580-603.
203. Stein, A. and P. Aloy, *Contextual specificity in peptide-mediated protein interactions.* PLoS One, 2008. **3**(7): p. e2524.
204. Chica, C., F. Diella, and T.J. Gibson, *Evidence for the concerted evolution between short linear protein motifs and their flanking regions.* PLoS One, 2009. **4**(7): p. e6052.
205. Mohan, A., et al., *Analysis of molecular recognition features (MoRFs).* J Mol Biol, 2006. **362**(5): p. 1043-59.
206. Neduva, V. and R.B. Russell, *Peptides mediating interaction networks: new leads at last.* Curr Opin Biotechnol, 2006. **17**(5): p. 465-71.
207. Landry, C.R., et al., *Extracting insight from noisy cellular networks.* Cell, 2013. **155**(5): p. 983-9.
208. Liu, B.A., B.W. Engelmann, and P.D. Nash, *High-throughput analysis of peptide-binding modules.* Proteomics, 2012. **12**(10): p. 1527-46.
209. Murray, E., et al., *Microarray-formatted clinical biomarker assay development using peptide aptamers to anterior gradient-2.* Biochemistry, 2007. **46**(48): p. 13742-51.
210. Lee, J.J., Y.S. Park, and K.J. Lee, *Hydrogen-deuterium exchange mass spectrometry for determining protein structural changes in drug discovery.* Arch Pharm Res, 2015. **38**(10): p. 1737-45.
211. Hoofnagle, A.N., K.A. Resing, and N.G. Ahn, *Protein analysis by hydrogen exchange mass spectrometry.* Annu Rev Biophys Biomol Struct, 2003. **32**: p. 1-25.
212. Pathan, M., et al., *FunRich: An open access standalone functional enrichment and interaction network analysis tool.* Proteomics, 2015. **15**(15): p. 2597-601.
213. O'Neill, R., *Personal communication, Thesis*. University of Edinburgh.
214. Schnell, U., J. Kuipers, and B.N. Giepmans, *EpCAM proteolysis: new fragments with distinct functions?* Biosci Rep, 2013. **33**(2): p. e00030.
215. van der Gun, B.T., et al., *EpCAM in carcinogenesis: the good, the bad or the ugly.* Carcinogenesis, 2010. **31**(11): p. 1913-21.
216. Stoecklein, N.H., et al., *Ep-CAM expression in squamous cell carcinoma of the esophagus: a potential therapeutic target and prognostic marker.* BMC Cancer, 2006. **6**: p. 165.

217. Linke, R., A. Klein, and D. Seimetz, *Catumaxomab: clinical development and future directions.* MAbs, 2010. **2**(2): p. 129-36.
218. Ogawa, K., et al., *EpCAM-targeted therapy for human hepatocellular carcinoma.* Ann Surg Oncol, 2014. **21**(4): p. 1314-22.
219. Fath, S., et al., *Multiparameter RNA and codon optimization: a standardized tool to assess and enhance autologous mammalian gene expression.* PLoS One, 2011. **6**(3): p. e17596.
220. Maslon, M.M., et al., *A divergent substrate-binding loop within the pro-oncogenic protein anterior gradient-2 forms a docking site for Reptin.* Journal of molecular biology, 2010. **404**(3): p. 418-38.
221. Healy, A., et al., *Discovery of a novel ligand that modulates theprotein-protein interactions of the AAA+ superfamily oncoprotein reptin.* Chemical Science, 2015. **epub ahead of print**.
222. Yu, J. and G.P. Smith, *Affinity maturation of phage-displayed peptide ligands.* Methods Enzymol, 1996. **267**: p. 3-27.
223. Bottger, V., et al., *Identification of novel mdm2 binding peptides by phage display.* Oncogene, 1996. **13**(10): p. 2141-7.
224. Vassilev, L.T., et al., *In vivo activation of the p53 pathway by small-molecule antagonists of MDM2.* Science, 2004. **303**(5659): p. 844-8.
225. Ray-Coquard, I., et al., *Effect of the MDM2 antagonist RG7112 on the P53 pathway in patients with MDM2-amplified, well-differentiated or dedifferentiated liposarcoma: an exploratory proof-of-mechanism study.* Lancet Oncol, 2012. **13**(11): p. 1133-40.
226. Dornan, D., et al., *The proline repeat domain of p53 binds directly to the transcriptional coactivator p300 and allosterically controls DNA-dependent acetylation of p53.* Mol Cell Biol, 2003. **23**(23): p. 8846-61.
227. Pauli, C., et al., *Tumor-specific glycosylation of the carcinoma-associated epithelial cell adhesion molecule EpCAM in head and neck carcinomas.* Cancer Lett, 2003. **193**(1): p. 25-32.
228. Munz, M., et al., *Glycosylation is crucial for stability of tumour and cancer stem cell antigen EpCAM.* Front Biosci, 2008. **13**: p. 5195-201.
229. Maetzel, D., et al., *Nuclear signalling by tumour-associated antigen EpCAM.* Nat Cell Biol, 2009. **11**(2): p. 162-71.
230. Dolle, L., et al., *EpCAM and the biology of hepatic stem/progenitor cells.* Am J Physiol Gastrointest Liver Physiol, 2015. **308**(4): p. G233-50.
231. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2005. **33**(Database issue): p. D514-7.
232. Stenson, P.D., et al., *The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.* Hum Genet, 2014. **133**(1): p. 1-9.
233. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
234. Vidal, M., M.E. Cusick, and A.L. Barabasi, *Interactome networks and human disease.* Cell, 2011. **144**(6): p. 986-98.

235. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
236. Yanagida, M., *Functional proteomics; current achievements.* J Chromatogr B Analyt Technol Biomed Life Sci, 2002. **771**(1-2): p. 89-106.
237. Mullard, A., *Protein-protein interaction inhibitors get into the groove.* Nat Rev Drug Discov, 2012. **11**(3): p. 173-5.
238. Dunker, A.K., et al., *Flexible nets. The roles of intrinsic disorder in protein interaction networks.* FEBS J, 2005. **272**(20): p. 5129-48.
239. Sarmady, M., W. Dampier, and A. Tozeren, *HIV protein sequence hotspots for crosstalk with host hub proteins.* PLoS One, 2011. **6**(8): p. e23293.
240. van den Heuvel, R.H. and A.J. Heck, *Native protein mass spectrometry: from intact oligomers to functional machineries.* Curr Opin Chem Biol, 2004. **8**(5): p. 519-26.
241. Rao, V.S., et al., *Protein-protein interaction detection: methods and analysis.* Int J Proteomics, 2014. **2014**: p. 147648.
242. Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4569-74.
243. Gazdar, A.F., et al., *Lung cancer cell lines as tools for biomedical discovery and research.* J Natl Cancer Inst, 2010. **102**(17): p. 1310-21.
244. Boonstra, J.J., et al., *Verification and unmasking of widely used human esophageal adenocarcinoma cell lines.* J Natl Cancer Inst, 2010. **102**(4): p. 271-4.
245. Lim, L.Y., et al., *Mutant p53 mediates survival of breast cancer cells.* Br J Cancer, 2009. **101**(9): p. 1606-12.
246. Fredriksson, S., et al., *Protein detection using proximity-dependent DNA ligation assays.* Nat Biotechnol, 2002. **20**(5): p. 473-7.
247. Soderberg, O., et al., *Direct observation of individual endogenous protein complexes in situ by proximity ligation.* Nat Methods, 2006. **3**(12): p. 995-1000.
248. Holden, S.J., et al., *Defining the limits of single-molecule FRET resolution in TIRF microscopy.* Biophys J, 2010. **99**(9): p. 3102-11.
249. Troester, M.A., et al., *Gene expression patterns associated with p53 status in breast cancer.* BMC Cancer, 2006. **6**: p. 276.
250. Lacroix, M. and G. Leclercq, *Relevance of breast cancer cell lines as models for breast tumours: an update.* Breast Cancer Res Treat, 2004. **83**(3): p. 249-89.
251. Stadler, C., et al., *Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells.* Nat Methods, 2013. **10**(4): p. 315-23.
252. Schnell, U., V. Cirulli, and B.N. Giepmans, *EpCAM: structure and function in health and disease.* Biochim Biophys Acta, 2013. **1828**(8): p. 1989-2001.
253. Voduc, D., C. Kenney, and T.O. Nielsen, *Tissue microarrays in clinical oncology.* Semin Radiat Oncol, 2008. **18**(2): p. 89-97.
254. Jawhar, N.M., *Tissue Microarray: A rapidly evolving diagnostic and research tool.* Ann Saudi Med, 2009. **29**(2): p. 123-7.
255. Allred, D.C., et al., *Prognostic and predictive factors in breast cancer by immunohistochemical analysis.* Mod Pathol, 1998. **11**(2): p. 155-68.

256. Braakman, I. and N.J. Bulleid, *Protein folding and modification in the mammalian endoplasmic reticulum.* Annu Rev Biochem, 2011. **80**: p. 71-99.
257. Wilson, S.M. and A. Bacic, *Preparation of plant cells for transmission electron microscopy to optimize immunogold labeling of carbohydrate and protein epitopes.* Nat Protoc, 2012. **7**(9): p. 1716-27.
258. Ladwein, M., et al., *The cell-cell adhesion molecule EpCAM interacts directly with the tight junction protein claudin-7.* Exp Cell Res, 2005. **309**(2): p. 345-57.
259. Huang, L., et al., *Potential pitfalls and solutions for use of fluorescent fusion proteins to study the lysosome.* PLoS One, 2014. **9**(2): p. e88893.
260. Argyro, F., et al., *The Anterior Gradient-2 Pathway as a Model for Developing Peptide-Aptamer Anti-Cancer Drug Leads that Stimulate p53 Function.* Current Chemical Biology, 2009. **3**(2): p. 124-137.
261. Frampton, J.E., *Catumaxomab: in malignant ascites.* Drugs, 2012. **72**(10): p. 1399-410.
262. Anelli, T. and R. Sitia, *Protein quality control in the early secretory pathway.* EMBO J, 2008. **27**(2): p. 315-27.
263. Griffiths, A.J.F., *An introduction to genetic analysis*. 7th ed. 2000, New York: W.H. Freeman. xvii, 860 p.
264. Beutler, B., *Immunology, phenotype first. Preface.* Curr Top Microbiol Immunol, 2008. **321**: p. v-viii.
265. Rouet, P., F. Smih, and M. Jasin, *Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease.* Mol Cell Biol, 1994. **14**(12): p. 8096-106.
266. Cox, D.B., R.J. Platt, and F. Zhang, *Therapeutic genome editing: prospects and challenges.* Nat Med, 2015. **21**(2): p. 121-31.
267. Hentze, M.W. and A.E. Kulozik, *A perfect message: RNA surveillance and nonsense-mediated decay.* Cell, 1999. **96**(3): p. 307-10.
268. Hsu, P.D., E.S. Lander, and F. Zhang, *Development and applications of CRISPR-Cas9 for genome engineering.* Cell, 2014. **157**(6): p. 1262-78.
269. Genovese, P., et al., *Targeted genome editing in human repopulating haematopoietic stem cells.* Nature, 2014. **510**(7504): p. 235-40.
270. Makarova, K.S., et al., *An updated evolutionary classification of CRISPR-Cas systems.* Nat Rev Microbiol, 2015. **13**(11): p. 722-36.
271. Ran, F.A., et al., *Genome engineering using the CRISPR-Cas9 system.* Nat Protoc, 2013. **8**(11): p. 2281-308.
272. Thompson, A., et al., *Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS.* Anal Chem, 2003. **75**(8): p. 1895-904.
273. Ross, P.L., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.* Mol Cell Proteomics, 2004. **3**(12): p. 1154-69.
274. Barrera, N.P. and C.V. Robinson, *Advances in the mass spectrometry of membrane proteins: from individual proteins to intact complexes.* Annu Rev Biochem, 2011. **80**: p. 247-71.
275. Savas, J.N., et al., *Mass spectrometry accelerates membrane protein analysis.* Trends Biochem Sci, 2011. **36**(7): p. 388-96.

276. Liu, H., et al., *Mass spectrometry-based analysis of glycoproteins and its clinical applications in cancer biomarker discovery.* Clin Proteomics, 2014. **11**(1): p. 14.
277. Stewart, G.D., et al., *The dermcidin gene in cancer: role in cachexia, carcinogenesis and tumour cell survival.* Curr Opin Clin Nutr Metab Care, 2008. **11**(3): p. 208-13.
278. Brauer, H.A., et al., *Dermcidin expression is associated with disease progression and survival among breast cancer patients.* Breast Cancer Res Treat, 2014. **144**(2): p. 299-306.
279. Porter, D., et al., *A neural survival factor is a candidate oncogene in breast cancer.* Proc Natl Acad Sci U S A, 2003. **100**(19): p. 10931-6.
280. Schittek, B., *The multiple facets of dermcidin in cell survival and host defense.* J Innate Immun, 2012. **4**(4): p. 349-60.
281. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-2.
282. Yu, H., et al., *Genomic analysis of essentiality within protein networks.* Trends Genet, 2004. **20**(6): p. 227-31.
283. Hahn, M.W. and A.D. Kern, *Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks.* Mol Biol Evol, 2005. **22**(4): p. 803-6.
284. Cho, S.W., et al., *Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease.* Nat Biotechnol, 2013. **31**(3): p. 230-2.
285. Doench, J.G., et al., *Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation.* Nat Biotechnol, 2014. **32**(12): p. 1262-7.
286. Wang, X., et al., *Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors.* Nat Biotechnol, 2015. **33**(2): p. 175-8.
287. Kim, D., et al., *Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells.* Nat Biotechnol, 2016. **34**(8): p. 863-8.
288. Abudayyeh, O.O., et al., *C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector.* Science, 2016. **353**(6299): p. aaf5573.
289. Gupta, A., et al., *Loss of anterior gradient 2 (Agr2) expression results in hyperplasia and defective lineage maturation in the murine stomach.* J Biol Chem, 2013. **288**(6): p. 4321-33.
290. Kappel, S., et al., *Silencing of mammalian genes by tetracycline-inducible shRNA expression.* Nat Protoc, 2007. **2**(12): p. 3257-69.
291. Ong, S.E. and M. Mann, *Mass spectrometry-based proteomics turns quantitative.* Nat Chem Biol, 2005. **1**(5): p. 252-62.
292. Ong, S.E., et al., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.* Mol Cell Proteomics, 2002. **1**(5): p. 376-86.
293. Ong, S.E. and M. Mann, *A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC).* Nat Protoc, 2006. **1**(6): p. 2650-60.
294. Bouchal, P., et al., *Biomarker discovery in low-grade breast cancer using isobaric stable isotope tags and two-dimensional liquid chromatography-tandem mass spectrometry (iTRAQ-2DLC-MS/MS) based quantitative proteomic analysis.* J Proteome Res, 2009. **8**(1): p. 362-73.

295. Peracaula, R., et al., *Altered glycosylation pattern allows the distinction between prostate-specific antigen (PSA) from normal and tumor origins.* Glycobiology, 2003. **13**(6): p. 457-70.
296. Gomaa, A.I., et al., *Diagnosis of hepatocellular carcinoma.* World J Gastroenterol, 2009. **15**(11): p. 1301-14.
297. Gray, T.A., et al., *Engineering a synthetic cell panel to identify signalling components reprogrammed by the cell growth regulator anterior gradient-2.* Mol Biosyst, 2014. **10**(6): p. 1409-25.
298. Luzio, J.P., P.R. Pryor, and N.A. Bright, *Lysosomes: fusion and function.* Nat Rev Mol Cell Biol, 2007. **8**(8): p. 622-32.
299. Bunton-Stasyshyn, R.K., et al., *SOD1 Function and Its Implications for Amyotrophic Lateral Sclerosis Pathology: New and Renascent Themes.* Neuroscientist, 2015. **21**(5): p. 519-29.
300. Way, L., et al., *Rearrangement of mitochondrial pyruvate dehydrogenase subunit dihydrolipoamide dehydrogenase protein-protein interactions by the MDM2 ligand nutlin-3.* Proteomics, 2016. **16**(17): p. 2327-44.
301. Song, C., et al., *Crystal structure and functional mechanism of a human antimicrobial membrane channel.* Proc Natl Acad Sci U S A, 2013. **110**(12): p. 4586-91.
302. Azhar, A., et al., *Recent updates on molecular genetic engineering approaches and applications of human therapeutic proteins.* Curr Protein Pept Sci, 2016.
303. Lipman, N.S., et al., *Monoclonal versus polyclonal antibodies: distinguishing characteristics, applications, and information resources.* ILAR J, 2005. **46**(3): p. 258-68.
304. Holliger, P. and P.J. Hudson, *Engineered antibody fragments and the rise of single domains.* Nat Biotechnol, 2005. **23**(9): p. 1126-36.
305. Cohen, S. and C. Milstein, *Structure of antibody molecules.* Nature, 1967. **214**(5087): p. 449-52 passim.
306. Porter, R.R., *The hydrolysis of rabbit y-globulin and antibodies with crystalline papain.* Biochem J, 1959. **73**: p. 119-26.
307. Nisonoff, A., et al., *Separation of univalent fragments from the bivalent rabbit antibody molecule by reduction of disulfide bonds.* Arch Biochem Biophys, 1960. **89**: p. 230-44.
308. Morrison, S.L., et al., *Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains.* Proc Natl Acad Sci U S A, 1984. **81**(21): p. 6851-5.
309. Ahmad, Z.A., et al., *scFv antibody: principles and clinical application.* Clin Dev Immunol, 2012. **2012**: p. 980250.
310. Bagshawe, K.D., S.K. Sharma, and R.H. Begent, *Antibody-directed enzyme prodrug therapy (ADEPT) for cancer.* Expert Opin Biol Ther, 2004. **4**(11): p. 1777-89.
311. Smith, G.P., *Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface.* Science, 1985. **228**(4705): p. 1315-7.
312. McCafferty, J., et al., *Phage antibodies: filamentous phage displaying antibody variable domains.* Nature, 1990. **348**(6301): p. 552-4.
313. Carmen, S. and L. Jermutus, *Concepts in antibody phage display.* Brief Funct Genomic Proteomic, 2002. **1**(2): p. 189-203.

314. Hoogenboom, H.R., *Overview of antibody phage-display technology and its applications.* Methods Mol Biol, 2002. **178**: p. 1-37.
315. Marks, J.D., et al., *By-passing immunization. Human antibodies from V-gene libraries displayed on phage.* J Mol Biol, 1991. **222**(3): p. 581-97.
316. Nelson, B. and S.S. Sidhu, *Synthetic antibody libraries.* Methods Mol Biol, 2012. **899**: p. 27-41.
317. Chothia, C., et al., *Conformations of immunoglobulin hypervariable regions.* Nature, 1989. **342**(6252): p. 877-83.
318. Huston, J.S., et al., *Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in Escherichia coli.* Proc Natl Acad Sci U S A, 1988. **85**(16): p. 5879-83.
319. Bird, R.E., et al., *Single-chain antigen-binding proteins.* Science, 1988. **242**(4877): p. 423-6.
320. Desplancq, D., et al., *Multimerization behaviour of single chain Fv variants for the tumour-binding antibody B72.3.* Protein Eng, 1994. **7**(8): p. 1027-33.
321. Merk, H., et al., *Cell-free expression of two single-chain monoclonal antibodies against lysozyme: effect of domain arrangement on the expression.* J Biochem, 1999. **125**(2): p. 328-33.
322. Huston, J.S., et al., *Protein engineering of single-chain Fv analogs and fusion proteins.* Methods Enzymol, 1991. **203**: p. 46-88.
323. Whitlow, M., et al., *An improved linker for single-chain Fv with reduced aggregation and enhanced proteolytic stability.* Protein Eng, 1993. **6**(8): p. 989-95.
324. Fitch, J.C., et al., *Pharmacology and biological efficacy of a recombinant, humanized, single-chain antibody C5 complement inhibitor in patients undergoing coronary artery bypass graft surgery with cardiopulmonary bypass.* Circulation, 1999. **100**(25): p. 2499-506.
325. Teeling, J.L., et al., *Characterization of new human CD20 monoclonal antibodies with potent cytolytic activity against non-Hodgkin lymphomas.* Blood, 2004. **104**(6): p. 1793-800.
326. Linsley, P.S., *New look at an old costimulator.* Nat Immunol, 2005. **6**(3): p. 231-2.
327. Goel, A., et al., *Genetically engineered tetravalent single-chain Fv of the pancarcinoma monoclonal antibody CC49: improved biodistribution and potential for therapeutic application.* Cancer Res, 2000. **60**(24): p. 6964-71.
328. May, C., P. Sapra, and H.P. Gerber, *Advances in bispecific biotherapeutics for the treatment of cancer.* Biochem Pharmacol, 2012. **84**(9): p. 1105-12.
329. Hanes, J. and A. Pluckthun, *In vitro selection and evolution of functional proteins by using ribosome display.* Proc Natl Acad Sci U S A, 1997. **94**(10): p. 4937-42.
330. Francisco, J.A., et al., *Production and fluorescence-activated cell sorting of Escherichia coli expressing a functional antibody fragment on the external surface.* Proc Natl Acad Sci U S A, 1993. **90**(22): p. 10444-8.
331. Boder, E.T. and K.D. Wittrup, *Yeast surface display for screening combinatorial polypeptide libraries.* Nat Biotechnol, 1997. **15**(6): p. 553-7.
332. Jain, S., et al., *The Development of a Recombinant scFv Monoclonal Antibody Targeting Canine CD20 for Use in Comparative Medicine.* PLoS One, 2016. **11**(2): p. e0148366.

333. Lee, B.R. and T. Kamitani, *Improved immunodetection of endogenous alpha-synuclein.* PLoS One, 2011. **6**(8): p. e23939.
334. Tsirigos, K.D., et al., *The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides.* Nucleic Acids Res, 2015. **43**(W1): p. W401-7.
335. Kohler, G. and C. Milstein, *Continuous cultures of fused cells secreting antibody of predefined specificity.* Nature, 1975. **256**(5517): p. 495-7.
336. Tutt, A.L., et al., *Monoclonal antibody therapy of B cell lymphoma: signaling activity on tumor cells appears more important than recruitment of effectors.* J Immunol, 1998. **161**(6): p. 3176-85.
337. Taylor, R.P., *Of mice and mechanisms: identifying the role of complement in monoclonal antibody-based immunotherapy.* Haematologica, 2006. **91**(2): p. 146a.
338. Clynes, R., et al., *Fc receptors are required in passive and active immunity to melanoma.* Proc Natl Acad Sci U S A, 1998. **95**(2): p. 652-6.
339. Weiner, G.J., *Building better monoclonal antibody-based therapeutics.* Nat Rev Cancer, 2015. **15**(6): p. 361-70.
340. Wu, A.M., et al., *Tumor localization of anti-CEA single-chain Fvs: improved targeting by non-covalent dimers.* Immunotechnology, 1996. **2**(1): p. 21-36.
341. Roberge, M., et al., *Construction and optimization of a CC49-based scFv-beta-lactamase fusion protein for ADEPT.* Protein Eng Des Sel, 2006. **19**(4): p. 141-5.
342. Kreitman, R.J., *Immunotoxins for targeted cancer therapy.* AAPS J, 2006. **8**(3): p. E532-51.
343. Falco, C.N., et al., *scFv-based fluorogen activating proteins and variable domain inhibitors as fluorescent biosensor platforms.* Biotechnol J, 2009. **4**(9): p. 1328-36.
344. Kontopidis, G., et al., *Insights into cyclin groove recognition: complex crystal structures and inhibitor design through ligand exchange.* Structure, 2003. **11**(12): p. 1537-46.
345. Plescia, J., et al., *Rational design of shepherdin, a novel anticancer agent.* Cancer Cell, 2005. **7**(5): p. 457-68.
346. Smirnov, D.A., et al., *Global gene expression profiling of circulating tumor cells.* Cancer Res, 2005. **65**(12): p. 4993-7.
347. Hapangama, D.K., et al., *Aberrant expression of metastasis-inducing proteins in ectopic and matched eutopic endometrium of women with endometriosis: implications for the pathogenesis of endometriosis.* Hum Reprod, 2012. **27**(2): p. 394-407.
348. Koegl, M. and P. Uetz, *Improving yeast two-hybrid screening systems.* Brief Funct Genomic Proteomic, 2007. **6**(4): p. 302-12.